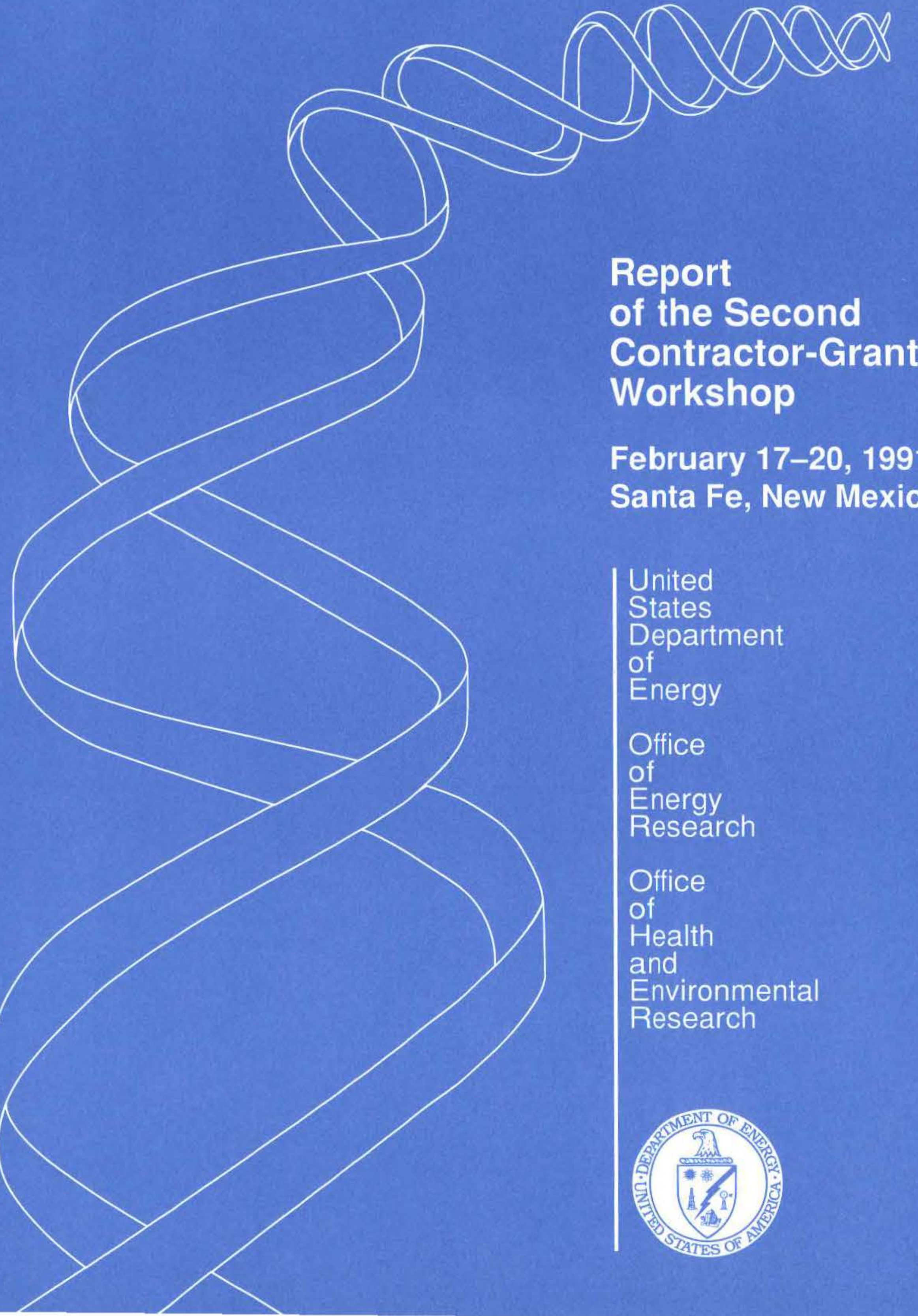
Conf-9102129

DOE

Human Genome Program



Report of the Second Contractor-Grantee Workshop

February 17–20, 1991
Santa Fe, New Mexico

United
States
Department
of
Energy

Office
of
Energy
Research

Office
of
Health
and
Environmental
Research

Acknowledgements

The information for this report was compiled and prepared for publication by Laura Yust and other Human Genome Management Information System staff members. Charles Cantor contributed the "Santa Fe Workshop Summary" with assistance from Elbert Branscomb, Anthony Carrano, Leroy Hood, Robert Moyzis, Robert Robbins, and DOE program staff. Sylvia Spengler (workshop organizer) and Eileen Mendez at the Lawrence Berkeley Laboratory Human Genome Center provided the abstracts. Most importantly, the contributions of grantees and contractors to the DOE Human Genome Program are gratefully acknowledged.

For information or additional copies of this report contact:

Laura Yust
Human Genome Management Information System
Oak Ridge National Laboratory
P.O. Box 2008
Oak Ridge, TN 37831-6050

615/574-7582, FTS 624-7582; Fax: 615/574-9888, FTS 624-9888;
Internet: "yustln@ornl.gov"; BITNET: "yustln@ornlstc.bitnet"

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

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

Price: Printed Copy A06
Microfiche A01

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

Conf-9102129
Dist. Category UC-408

DOE Human Genome Program

Report of the Second Contractor-Grantee Workshop

February 17–20, 1991
Santa Fe, New Mexico

Date Published: August 1991

Prepared for the
U.S. Department of Energy
Office of Energy Research
Office of Health and Environmental Research
Washington, D.C. 20585
under budget and reporting code KP 0404000

Prepared by
Human Genome Management Information System
Oak Ridge National Laboratory
Oak Ridge, TN 37831-6050

Managed by
MARTIN MARIETTA ENERGY SYSTEMS, INC.
for the
U.S. DEPARTMENT OF ENERGY
UNDER CONTRACT DE-AC05-84OR21400

Preface

For over four decades the Office of Health and Environmental Research of the Department of Energy (DOE) and its predecessor organizations have supported major research on the genetic effects of energy-related agents. Although knowledge and capabilities in this area have advanced in many ways, DOE realized in the early 1980s that more complete genetic resources would be required for tasks such as measuring mutations in descendants of the Hiroshima and Nagasaki survivors. These novel resources would include physical maps of the 24 different human chromosomes, a complete human genome DNA base sequence, and the computational tools to use this information. Partly because of the need for new and different approaches to modern challenges, DOE began the Human Genome Initiative in 1987.

The number of genome-funded projects has increased rapidly, and the initiative has grown into a flourishing program, with a complementary counterpart at the National Institutes of Health. The impact of the Human Genome Project on science and society will be substantial as significant achievements become frequent occurrences. New physical mapping strategies have allowed our national laboratories and university grantees to move ahead in a timely manner, and development of innovative approaches to determining the base sequence of chromosomes is proceeding at an encouraging rate. Our computational capabilities are improving. New networking workstations and methods of inputting and analyzing mapping and sequencing data are being developed to allow investigators easy access to data from major databanks. Investigations are under way into the ethical, legal, and social implications of the use of data generated by this program. The first of many DOE Human Genome Distinguished Postdoctoral Fellowships have been awarded to train additional scientists in areas critical to the success of the Human Genome Program.

Program management needs, such as project coordination and the desire to create an environment in which the program's investigators could interact, formed the basis for this workshop. We are indebted to Sylvia Spengler and the session chairs, but especially to all contributors.

The 1991 workshop had many stimulating presentations, and numerous collaborative efforts have developed, but we anticipate that its successor, scheduled for the fall of 1992, will provide additional concepts, strategies, and technologies that are not yet imagined and that are based on multidisciplinary interactions. We await with excitement what lies ahead.

Benjamin J. Barnhart
Human Genome Program
Office of Health and Environmental
Research
DOE Office of Energy Research

Table of Contents

*The abstracts were reproduced as closely as possible to the author's original.

Workshop Agenda

Sunday, February 17

12 noon – 5:30 pm: Registration
5:00 – 7:30 pm: Mixer, video talk by Senator Domenici

Monday, February 18

8:30 am – 12 noon: Chair, C. Cantor

Introduction: B. Barnhart
Human Genome Program in the Overall DOE Context: D. Galas

Mapping: Progress and New Methods:
G. Evans
R. Stallings
G. Sutherland
T. Caskey
M. Simon
P. de Jong

Lunch Talk on Ethical, Legal, and Social Issues: M. Yesley

2:00 – 5:30 pm: Chair, R. Moyzis

Cloning/Amplification:
V. Rao
S. Antonarakis
M. McCormick
P. Youderian
D. Nelson
G. Hermanson

5:30 – 7:00 pm: Poster Session

Tuesday, February 19

8:30 am – 12 noon: Chair, R. Robbins

Informatics: New Methods, Unresolved Problems:
T. Slezak
J. Fickett
R. Douthart
D. Searls
T. Hunkapiller
E. Lawler

Tuesday, February 19 (continued)

3:30 – 6:00 pm: Chair, E. Branscomb

Database Needs for Sequencing Projects:

R. Gesteland
C. Lawrence
L. Hood

Public Databases:

P. Pearson
C. Burks

7:00 – 9:00 pm: Buffet and Poster Session

Wednesday, February 20

8:30 am – 12 noon: Chair, A. Carrano

DNA Sequencing Methodology:

L. Smith
B. Karger
D. Berg
A. Riggs
R. Keller
F. W. Studier

2:00 – 5:30 pm: Chair, L. Hood

New Techniques and Instruments:

B. Trask
B. Jacobson
J. C. Giddings
R. Kopelman
R. Foote
R. Athwal
J. Vos

Meeting Summary and Perspective: C. Cantor

Santa Fe Workshop Summary*

The Department of Energy (DOE) Human Genome Program held its Second Contractor-Grantee Workshop in Santa Fe, New Mexico, on February 17–20, 1991. Almost all the DOE-supported projects were represented, either by oral presentations or by posters. More than 200 DOE-supported scientists attended the meeting, in addition to invited guests and representatives from a number of interested industries. Benjamin Barnhart, Manager, Human Genome Program, Office of Health and Environmental Research (OHER), welcomed the attendees and gave a brief overview of research that would be presented as platform and poster sessions throughout the 3-day workshop.

Six platform sessions focused on database and computer algorithm needs for existing or projected genome research, large DNA fragment cloning, progress in physical mapping, DNA sequencing instrumentation, strategies for preparing samples for efficient DNA sequencing, and new methods for a variety of genome efforts. David Galas, Associate Director, OHER, spoke about the relationship between the Human Genome Program and other OHER research programs. He emphasized the importance of new strategies, resources, and technologies being developed in the Human Genome Program for research aimed at understanding genomic damage by radiation and chemicals. He also stated that DOE is uniquely positioned to make major advances in protein structure and function relationships, because DOE's extensive capabilities in structural biology at the national laboratories will be coordinated with those of the Human Genome Project. Michael Yesley [Los Alamos National Laboratory (LANL)] presented ethical, legal, and social issues pertaining to data produced in the genome project.

The general impression conveyed by most presenters was that physical mapping efforts are going well; chromosomes 16, 19, and portions of 11 are now well covered with large numbers of assembled contigs. Fluorescence in situ hybridization (FISH) is emerging as an extremely effective method for ordering cloned probes.

Automated DNA sequencing is becoming more efficient through approaches like capillary or thin gel electrophoresis. Informatics support for most current physical mapping efforts and for large-scale DNA sequencing needs to be increased, and more focus is needed on ongoing biology projects. At present there are many parallel efforts in genome technology, informatics, cloning, mapping, and sequencing. This is a healthy situation, but if the genome is to be mapped and sequenced at the rate set by the NIH-DOE 5-year plan and the DOE program plan, those efforts that prove most viable will need to be selectively encouraged. One frequently recurring theme during the 3-day meeting was the rapid change in bottlenecks or rate-limiting steps, particularly in DNA sequencing. In the near future, not only does close attention need to be paid to generation of sequence data, but also to speeding up production of DNA samples for sequencing and faster assembly and analysis of extended sequence data.

Physical Mapping

In general, impressive progress continues in the construction of physical maps of selected human chromosomes. Yeast artificial chromosomes (YACs) are proving helpful in linking cosmid contigs or in providing regional coverage more rapidly than cosmids or smaller clones. FISH is emerging as a powerful technique at several levels of resolution. With metaphase chromosomes, FISH offers a rapid and accurate method of regional clone assignment to a chromosome band. Barbara Trask [Lawrence Livermore National Laboratory (LLNL)] showed that, with interphase cells, much higher resolution mapping can be carried out; once markers are known to be close, their relative order can be determined and distances estimated in the 100-kb to 1-Mb range. Another powerful tool is a nested set of chromosomal deletions, particularly if these are moved into rodent cells. The presence of a selectable marker near the long-arm telomere of chromosome 16 made the construction of such a set of hybrid cell lines particularly convenient for that chromosome, as shown by Grant Sutherland (Adelaide Children's Hospital, South Australia). He also showed the usefulness of FISH in characterizing chromosomal rearrangements.

Working on Chromosome 11, Glen Evans (Salk Institute) and his collaborators have combined a series of mapping methods to construct several multimegabase-sized contigs of cloned DNA segments. The methods include FISH, pulsed-field gel electrophoresis, and rapid walking using probes generated from the ends of shorter contigs. Instead of going through the laborious task of subcloning YACs into cosmids for subsequent high-resolution analyses, Evans described the use of a simple procedure in which a YAC serves as a hybridization probe against a filter array of cosmids to identify those that correspond. Overlapping YACs can be detected by the partially overlapping pattern of cosmids. This method appears to have considerable promise and generality. With the appropriate libraries, a series of overlapping clones for 1- to 2-Mb regions can apparently be isolated in a matter of weeks.

Both the chromosome 19 program at LLNL and that of chromosome 16 at LANL have collected about two-thirds of their chromosomes into cosmid contigs. A variety of effective, automated approaches are being used to expand and link these contigs. Both mapping projects are reaching the end game, but how long the end game will take and what a finished map should be are unclear. Existing maps are dense enough to be quite useful to those who wish to locate genes on these chromosomes; about 80% of randomly picked probes will fall on contigs or clones already mapped. To aid such studies, a series of anchor points have been established between the current genetic and physical maps. No clone-order discrepancies are evident between the two types of map for either chromosome. Distortions in relative distances on the two maps result from variations in the frequency of meiotic recombination along the chromosome.

Excellent progress is being made with maps of several X-chromosome regions by Thomas Caskey, David Nelson (Baylor College of Medicine), and their coworkers, who are focusing on several areas that contain disease genes of interest rather than attempting to construct a complete map of this very large chromosome. Caskey described the enormous power of tandem simple sequence repeats used as genetic markers. In one X-chromosome region, an informative marker could be found every 80 kb. The polymerase chain reaction (PCR) using bubble primers is proving effective in locating single-copy sequences adjacent to these repeats.

In the next few years, complete contig maps should be obtained for chromosomes 11, 16, 19, and X. While little useful data are produced in early physical mapping projects, map usefulness increases rapidly with regional assignment of contigs and the production of cosmid YAC filter arrays. Many DOE efforts are now at the point where much of their potential utility can be realized.

Large-Insert Cloning Vectors

A number of groups reported new approaches toward large-insert cloning vectors. Peter Hahn (State University of New York, Syracuse) and his colleagues employed double-minute chromosomes as megabase-cloning vehicles, an application that is interesting because the double minutes are relatively stable and can replicate megabase-size DNA inserts. A selectable marker gene is randomly inserted into chromosomes and then selectively used for chromosome fragmentation to multimegabase size, followed by amplification to generate appropriate drug resistance. This work looks intriguing but is in the earliest stages. Application of the technique for random amplification of a DNA segment, and purification of the double-minute chromosomes, may provide some interesting challenges.

A second approach was the use of Epstein-Barr virus (EBV) as a cloning vehicle. The premise is that human cells can tolerate large repetitive sequences and could serve as ideal vehicles for the amplification of large DNA inserts. Once again, work in this system is in the very early stages. Precautions must be taken to ensure the health of investigators working with EBV.

A third approach, reported by Hiroshi Shizuya from Melvin Simon's (California Institute of Technology) laboratory, involved a new cloning system employing *Escherichia coli* and its plasmid F factor. This bacterial artificial chromosome vector has several limitations including the possibility of recombination, but it seems promising for exploration.

Philip Youderian (California Institute of Biological Research) reported on a set of "stealth" vectors that carry an S replicon, the *Salmonella* phage B-22 early region, and a chloramphenicol resistance determinant. The stealth vector S110 can accept several hundred kilobase inserts of DNA, maintain them in single copy, and then amplify them about 500-fold. These studies are also in the earliest stages, and several questions about stability and feasibility remain unanswered.

Gary Hermanson from Glen Evans' laboratory gave an interesting presentation on isolating the ends of YAC inserts by homologous recombination. This method promises to provide a powerful technique for walking within YAC libraries.

The extent of rearrangements is an unresolved concern in most of the above systems and others such as P1 and T4 bacteriophage cloning in *E. coli.*

Sherman Weissman (Yale University) reported on his group's effort to achieve normalized cDNA libraries from human thymus. These efforts appear to be proceeding very well and set the stage for a large-scale sequence analysis of corresponding cDNAs from various tissues.

DNA Sequencing Methodology

A goal of the Human Genome Project is to reduce DNA sequencing costs so major segments of the human genome can be analyzed. Automated DNA sequencers have dramatically improved the data-collection rate, but DNA sequencing is still not cost-effective for large-scale (megabase) efforts. Sequencing methodology can be divided into at least three components: template preparation for sequencing, data collection, and data analysis. While considerable attention has been given to data-collection instrumentation and data-analysis hardware and software, more emphasis is needed on technologies that will reduce sequencing costs by increasing throughput and template quality.

A number of presentations dealt with the use of transposon insertions to assist in DNA sequencing and minimize the redundancy associated with shotgun cloning strategies. Douglas Berg (Washington University School of Medicine) and his group have exploited the bacterial Tn5*supF* transposon system by inserting a known 260-bp sequence into lambda clones to serve as the priming site for sequencing. This procedure eliminates the need to subclone large inserts because sequencing could be accomplished directly from the λ clone by primer walking from the inserted transposon. Claire Berg (University of Connecticut) reported on the use of transposon γδ (Tn1000) to sequence plasmid inserts. Transposon-mediated sequencing offers the potential of scaling up to larger clone inserts such as cosmids; a disadvantage is that the transposon integration occurs at random, and clones having a distribution of integration sites must be selected for complete coverage of the insert. Robert Weiss and Raymond Gesteland (University of Utah) reported on the use of the γδ system to sequence plasmids containing 10-kb inserts. They developed a clone-pooling scheme to rapidly map the transposon-integration sites and then used the mapped transposons as primer sites for multiplexed dideoxy sequencing.

The concept of priming DNA sequencing reactions, using sequences within the insert itself, was presented by William Studier (Brookhaven National Laboratory). A library of nonamers or decamers would be established that could be annealed to the cosmid insert and serve as the initiation site for primer walking. To sequence any cosmid, the 260,000 possible nonamers could potentially be reduced to about 7000. Work is under way to test the efficiency of priming with these nonamers and to develop methods to join shorter primers into longer ones.

George Church (Harvard Medical School) presented the status of his developments for computer-assisted multiplexed sequencing. In this approach, 40 sequencing reaction sets, each tagged with specific oligonucleotides, are pooled in each lane (up to 60 lanes) of a gel. The DNA fragments are transferred to membranes, then hybridized with radiolabeled oligonucleotides complementary to the primer sets. Each filter can be reprobed up to 40 times, and 75 membranes can be handled simultaneously. The resulting films are automatically digitized by a scanning device, and an algorithm provides base calling. The software can do this analysis very well, although typically only 300 bases are analyzed. DNA sequencing with runs this short is likely to be more effective in prokaryotes than in eukaryotes with repetitive sequence elements. Some runs out to 800 to 900 bases have been achieved.

Arthur Riggs (Beckman Research Institute) described a ligation-mediated genomic sequencing strategy that allows sequence information to be derived from genomic DNA without subcloning. First a nested set of genomic fragments is created by Maxam-Gilbert cleavage reactions, and a known sequence-specific primer is annealed to the fragment mixture to identify the fragment of interest. The sequence-specific primer is used in an extension reaction to create a fragment with one blunt end to which a linker can be ligated. An exponential PCR reaction amplifies the fragment between the two priming sites, and the amplified fragments are run on a sequencing gel to create the sequence ladder. This technique is useful for sequences that are difficult to clone or for which directional sequencing is desirable. The methodology is also amenable to automation, but its applicability to large-scale sequencing remains to be demonstrated.

New Techniques for DNA Sequencing

Bruce Jacobson [Oak Ridge National Laboratory (ORNL)] and Heinrich Arlinghaus (Atom Sciences, Inc.) are exploring the use of tin, iron, and lanthanide isotopes to serve as reporter groups for DNA sequencing. Sputter-Initiated Resonance Ionization Spectroscopy and Laser Atomization Resonance Ionization Spectroscopy, two forms of mass spectrometry used as assays, offer the promise of striking sensitivity and speed of analysis. The thin-layer capillary electrophoresis techniques reported by Barry Karger (Northeastern University) present notable opportunities for speeding up throughput in the analysis of DNA sequence fragments. Lloyd Smith (University of Wisconsin) presented impressive data suggesting that thin-layer capillary gels will work effectively in the separation of sequencing fragments and provide multiple-lane parallelism for DNA sequence analysis. These techniques seem likely to be applied as the analytic system for most automated fluorescent sequencing devices now extant. Single DNA molecule analysis using exonuclease and single-nucleotide detection are being advanced by the Los Alamos group. Some progress has been made in the detection side of this automated fluorescent flow cytometry–based technology, but much work remains in enzymology, reporter labeling, and tethering the DNA molecules.

Many presentations were made on scanning tip microscopy, including scanning tunneling electron microscopy (STEM), atomic force microscopy, and scanning tunneling exciton microscopy. An exciting presentation by Rodney Balhorn and Wigburt Siekhaus (LLNL) described the use of STEM to reveal images of monolayers of adenine and thymine bases that are consistent with the known physical structure of these molecules.

Several other presenters described methods for detecting DNA for either sequencing or mapping. Christopher Martin and Irena Bronstein (Tropix, Inc.) demonstrated the application of chemiluminescence in Sanger dideoxy sequencing protocols, a technique that uses biotinylated primers in standard sequencing reactions. The gels are transferred to membranes that are later treated with streptavidin-alkaline phosphatase and the chemiluminescent substrate AMPPD. Film exposure time for chemiluminescent-signal detection was as short as 1 minute, thus reducing the DNA clone-sequencing time to about 8 to 9 hours.

Richard Mathies and Alexander Glazer (University of California) presented some elegant methods for high-sensitivity fluorescent DNA detection in gels. DNA is labeled with ethidium homodimer and/or thiazole orange, then loaded onto a gel for electrophoresis. After fragment separation, an argon ion laser is focused in the gel; then in turn the laser-excited fluorescence is focused on a confocal spatial filter, a spatial filter, and a photodetector. The gel is placed on a computer-controlled scan stage, and the scanned gel fluorescence image is stored and analyzed. Picogram sensitivity is achievable with the ethidium homodimer. The system has application to DNA sequencing and mapping gels.

Two novel methods are being developed as third-generation sequencing approaches. Joe Gray (LLNL) and coworkers are exploring X-ray diffraction for DNA sequence analysis. DNA to be sequenced is amplified, separated into four fractions, and each of the four bases is labeled with a heavy metal, such as I or Pt, that efficiently scatters X rays. After labeling, DNA molecules are aligned either by pulling DNA fibers or by preparing liquid crystalline DNA in the presence of a magnetic field. Fiber illumination with partially coherent X rays produces a scattering pattern that is deconvoluted by Fourier analysis, and sequence information is recorded by measuring the distances between labels in each of the four fractions. The approach has been successfully modeled on a theoretical basis and is about to enter proof-of-principle testing.

The second novel sequencing approach was presented by Radoje Drmanac and Radomir Crkvenjakov (Argonne National Laboratory). In sequencing by hybridization, short oligomers (8 to 9 mers) with known sequence are hybridized to DNA fragments with unknown sequence. The collection of oligomers that hybridize to a fragment define the sequence. With a 100-bp DNA segment of known sequence, a successful demonstration of core methodologies has been achieved. Procedure scale-ups and a second proof-of-concept test on DNAs of unknown sequence are in progress.

Informatics and Database Needs

When the Human Genome Project is finished, many of the innovative methodologies involved in its successful completion will remain in the specialists' domain or become part of the history of molecular biology. What will remain as the project's enduring contribution is a vast amount of computerized knowledge. Seen in this light, the Human Genome Project is an effort to create the most important database ever attempted—the database containing instructions for human development.

The 3.3 billion nucleotides in the DNA of a human gamete constitute a single set of these instructions. With each nucleotide represented as a single letter, one copy of this sequence typed in standard pica typeface on a continuous ribbon of material would reach from San Francisco to New York and then on to Mexico City. No unaided human mind could hope to comprehend this mass of information. Just storing such a sequence requires a computer if data is to be available in useful form. Representing individual variations and managing a fully annotated, functionally described version of the sequence is probably beyond current information-handling technology.

Even now, when the Human Genome Project is in the first year of its first 5-year plan, computer systems are playing an essential role in all phases of the work. Laboratory databases help manage research materials while computer-controlled robots perform experimental manipulations. Automated data-acquisition systems log experimental results, and analytical software assists in their interpretation. Local database systems store the accumulating knowledge of a research team, while public databases provide a new kind of publication for scientists to share their findings with the world.

Presentations at the Santa Fe meeting showed that progress is being made along a broad front in applying computer technology to the Human Genome Project. Jim Fickett (LANL) and colleagues provided an overview in which they noted that the growing map of chromosome 16 now requires the management of nearly a million data items, with the bulk coming from fingerprinting about 4000 clones. With 60% of chromosome 16 covered by cosmid contigs, attention is shifting to the use of YACs and sequence tagged sites (STSs) to close intercontig gaps, thereby requiring extensions to the existing data system. Collaborations with LLNL, GenBank®, Genome Data Base (GDB), and others are under way to facilitate development of improved tools and of improved methods for data exchange.

In preparation for an increasing role for STSs, Chris Fields (New Mexico State University) and colleagues described an automated system for screening candidate STS sequences and predicting good PCR priming sites on them. Sequences with similarity to known human repetitive elements are rejected, then the remaining sequences are compared with primate entries in GenBank and matching regions are flagged. Unflagged regions are analyzed and region pairs with similar C+G content and low secondary-structure probability are identified as potential priming sites. Finally, long, open reading frames (orfs) are translated and compared with entries in the Protein Information Resource (PIR). Three output files are generated for each candidate sequence: a summary of sub-sequence matches in GenBank and of predicted best priming sites, matches between long orfs and PIR entries, and an archival log of all operations performed.

Many posters and one presentation addressed the problem of sequence matching to genome computing. Eugene Lawler discussed a new dynamic programming method, developed with W. I. Chang (both at University of California, Berkeley), that is 4 or 10 times faster (using nucleotide or amino acid sequences, respectively) than the best previous dynamic programming algorithm. Lawler also discussed the "approximate substring matching problem," which can be stated as "given an integer, n, find the matches between the text and all fragments of length n of the pattern." At present, most attempts to retrieve database sequences based on similarities to another sequence require performing some sequence-analysis algorithm against every sequence in the database. Solutions to the substring-matching problem, whether involving suffix trees, matching statistics, hashing algorithms, or other methods, may allow indexed sequence retrievals based on their substring contents. If so, this would be of major importance for sequence databases.

Providing and controlling access to multiple data sources is rapidly becoming a key problem for genomic research, and permitting remote data access while meeting important security and integrity concerns will soon be essential. Thomas Slezak offered an insightful analysis, developed with Elbert Branscomb (both at LLNL), of this and other distributed database problems. Slezak reported on some proof-of-concept tests that showed how controlled, multiple-database access can be achieved by using the client-server architecture found in most commercial, multiuser database systems. If all participating sites use the same commercial product, relatively little programming effort is required. The report concludes that "multi-collaborator and distributed genome databases are technically feasible, necessary, and worth the overhead in data storage and database administration."

Another approach to managing local research-team data was presented by Manfred Zorn [Lawrence Berkeley Laboratory (LBL)] and others. Zorn described prototyping work for the comprehensive and general Chromosome Information System, which will become a major part of the computing effort at the LBL Human Genome Center. SuperCard on a Macintosh acts as the interface manager, with Sybase providing the database backend. Next the interface system will be ported to X-windows. The database has been designed at a level of abstraction close to the biologists' view of the information. ERDRAW (Entity-Relationship Schema Drawing Program) and SDT (Schema Design Tool), two database implementation tools also developed at LBL and available to interested parties, have facilitated the effort.

In time, public databases will become a major publishing outlet for many discoveries from the Human Genome Project. Peter Pearson (Johns Hopkins University) described in some detail the public database efforts under way with the GDB project. He also discussed ongoing collaborations with LANL and LLNL for sharing physical mapping data and for possible viewing by researchers of laboratory-generated map data in the information-rich context of GDB.

Many informatics posters offered a chance to become familiar with work in progress in a number of areas, from (1) data acquisition (robotics, image acquisition, and laboratory notebook projects) through preliminary data interpretation (contig assembly, Southern blots, fingerprints, and probed partial digests); (2) data analysis (pattern matching, sequence analysis, and sequence alignments); (3) local data management (reports from LANL, LBL, and LLNL); (4) software to facilitate access to public databases (GENCORE and Gnomeview); (5) database interfaces (control, security, and standards); and (6) tool development (logic programming and object manipulation). Betty Mansfield and her colleagues from ORNL presented a poster describing their efforts to develop the Human Genome Management Information System (HGMIS) to track the genome project itself. HGMIS also publishes DOE program and workshop reports and a bimonthly newsletter jointly sponsored by NIH and DOE. *Human Genome News* features technical articles, meeting reports, project news, genome events and training calendars, and funding announcements.

*Charles R. Cantor and HGMIS gratefully acknowledge contributions to this article by Elbert W. Branscomb, Anthony V. Carrano, Leroy E. Hood, Robert K. Moyzis, Robert J. Robbins, and DOE program staff.

Speaker Abstracts

Informatics: New Methods, Unresolved Problems

Cloning/Amplification

Mapping: Progress and New Methods

DNA Sequencing Instrumentation/Database Needs

DNA Sequencing Methodology

New Techniques and Instruments

Speaker Abstracts

Informatics: New Methods, Unresolved Problems

S01
HGIR: Information Management for a Growing Map

J. W. Fickett, H. T. Brown, C. Burks, M. J. Cinkosky, F. R. Fairfield, D. Nelson,
R. M. Pecherer, D. M. Sorensen, and R. D. Sutherland
Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos,
NM 87545

The growing map of human chromosome 16 now requires that the LANL project manage on the order of a million data items. Computer assistance is essential for acquisition, for well structured storage, for analysis, and for retrieval.

The bulk of this data comes from the fingerprinting of about 4000 clones (see abstract by Moyzis *et al.*). The fingerprint data is of two types; acquisition of each type is computer-assisted. Fragment lengths from restriction digests of each clone are determined on a BioImage Visage 110 workstation and transferred by network to the Laboratory Notebook (see abstract by Nelson and Fickett) database. Hybridization signals from Southern blots with repetitive DNA are determined in SCORE, a program we developed to aid the user in matching the digest gel image to the blot autoradiogram in order to interpret the latter (see abstract by Cannon *et al.*). The results from SCORE are also written to the database.

About 60% of chromosome 16 has now been covered by cosmid contigs. At present increasing emphasis is being placed on using YACs and STSs to close gaps between contigs, and on mapping contigs to the chromosome. Thus the Laboratory Notebook, whose first implementation supported primarily the fingerprinting process, has been expanded significantly to include sequence data and relative position information on all kinds of map elements. A simple and general data structure has been developed which will allow easy expansion of the database to new kinds of map data as the project progresses.

To allow easy access to all data for those without special computer training, an intuitive forms-based interface to the database has been developed and continues to grow. We are developing (in collaboration with GenBank®) an interface which is schema-driven. This will provide the same ease of access which is given by the custom forms, but at much lower software development/maintenance costs. It will also allow other laboratories to set up similar databases at low cost.

As the complexity of the map increases, the need for computer assistance in map assembly is becoming clear. A map editor is being developed which will allow the handling of all types of map data in a single environment.

Pooling of information from different groups will be a key factor in efficiently building a useful combined map. HGIR is collaborating with LLNL and GDB to make map data assembled at the national labs centrally available (see abstract by Nelson *et al.*). HGIR also continues to contribute to the development of multi-database access software.

All software developed as part of HGIR is freely available. Contact James Fickett at (505) 665-5340 (voice), at jwf@life.lanl.gov (electronic mail), or at the above address.

S02

Automated Prediction of Priming Sites for STS Sequences

C. A. Fields, B. Rappaport, C. A. Soderlund, V. Church,* C. E. Hildebrand,* and R. K. Moyzis*
Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003-0001
*Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545

An automated system for screening candidate STS sequences, and predicting good PCR priming sites on such sequences, is being developed to support STS mapping projects at the LANL Center for Human Genome Studies. The software, which is currently being developed and tested, performs the following functions for each candidate STS sequence.

1. Sequences and their complements are compared with consensus sequences of known human repetitive elements, using **lfasta**. Sequences with significant matches to a repeat consensus are rejected for use as STSs.

2. The remaining sequences and their complements are compared with all sequences in the primate section of GenBank® using **fasta**. Sequence regions having significant similarity to one or more sequences in GenBank are marked.

3. The unmarked regions of each sequence are analyzed for C+G content, ability to form secondary-structure hairpins, and complementarity to each other. Pairs of noncomplementary sequences regions with similar C+G content and low secondary-structure probability are identified as potential priming sites.

4. All sequences are scanned for long open-reading frames. Any long open-reading frames are translated, and the predicted amino-acid sequences are compared with the sequences in the PIR database using **fasta**. This procedure provides a useful first check for STS sequences that overlap coding exons.

The parameters controlling the sensitivities of the database searches, significance of matches, C+G content, secondary-structure probability, and length of open-reading frames to consider significant are set by the user. Three output files are generated for each candidate STS sequence, which contain: i) a summary of matches to sequences in GenBank, and the predicted best primer sites, ii) matches between translations of long orfs and protein sequences in the PIR, if any, and iii) an archival log of all operations performed.

S03

Approximate Pattern Matching and Biological Applications

E. L. Lawler and W. I. Chang
Computer Science Division, University of California, Berkeley, CA 94702

The proposed *sequence tagged sites* (STS) genomic map database of the Human Genome Project, as well as several mapping strategies, require the approximate matching of DNA sequences. Algorithms based on dynamic programming calculate most if not all entries of an *m* by *n* table, where *m*, *n* are respectively the lengths of the pattern and text sequences. We have done careful theoretical and empirical comparisons of these methods; apart from questions of overhead, they do not have the same dependence on alphabet size. Very recently we (joint work with J. Lampe) have discovered a speedup of the dynamic programming method whose running time depends on the *row averages* of the table (the higher the averages, the *faster* our algorithm). Our method works by efficiently partitioning each column into runs of consecutive integers, and is *four or ten times faster* (nucleotide and amino acid sequences, respectively) than the best previous algorithm based on dynamic programming. It has also given us special insights into the statistics of sequence matching.

At the 1990 *FOCS* and *HG II* conferences we posed and initiated the study of the *approximate substring matching problem*: Given an integer *l*, find all (approximate) matches between the text and length *l* fragments of the pattern. For example, when *l* is small this is a general method for finding *local similarities*. Variations and even more general models can be defined along these lines, and we have developed several fast algorithms based on *suffix trees* and *matching statistics* (a summary of all exact matches between fragments of the text and pattern). This approach appears to us to provide the proper framework for biological sequence analysis.

We have implemented many of our algorithms in the C programming language, and intend to make available a suite of sequence analysis tools.

Supported by DOE grant DE-FG03-90ER60999, July 1990–January 1991.

S04

Data Control and Security Issues for Distributed Genome Databases

Tom Slezak and Elbert Branscomb
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

Biological data has traditionally been stored in private databases, inaccessible to other researchers, or in various centralized repositories that provide dial-up access or update subscription services. It is becoming clear that such unrestricted private retention of large data sets generated with public funds should no longer be acceptable for genome grants. Centralized repository databases play many important roles for biological researchers, yet lack the immediacy of access to "fresh" data that is desired for collaborative research and community support. Recent advances in networking and databases permit a 3rd option: cooperating individual databases that function as a larger "distributed" database yet maintain complete local autonomy.

Some modern commercial relational databases provide true server/client access, which opens the possibility of allowing collaborators to have direct access to each other's data, under complete control of the owner. All details of access to remote database(s) can be hidden from end users with suitable front-ends. Issues of data security are in general handled trivially by the database package (by allowing read-only access to specified tables/views to external collaborators). Data control issues are more complex (e.g., if we collaborate with A and B by sending them clones/probes and put their results into our database, if both A and B are external users of our database they may not want the other to be able to view "their" data until they publish.) One suggested solution is to place a six-month hold on data before it could seen by non-collaborators.

Without arguing the merits or ethics of this approach, we have decided to demonstrate the technical feasibility of multi-collaborator and distributed genome databases. Experiments were conducted with the help of Debra Nelson of LANL that proved the ease of use of allowing controlled access to external databases over the Internet. Other experiments were done at LLNL to verify that the 6-month "data hold" concept was feasible, at the cost of adding owner and timestamp fields to any tables that might contain "proprietary" data.

We view this means of data sharing as an important tool for communities of tightly-coupled collaborating researchers willing to work with external databases at a fairly low level. We recognize that objections may be raised by sites that have databases from different vendors, and note that Sybase front-end tools are relatively inexpensive for genome researchers. In addition, the LLNL C-language interface library was designed to be portable to other relational databases and would make it easy to "import" data from a collaborator's "foreign" database, or to do cross-database pseudo-join queries.

We conclude that multi-collaborator and distributed genome databases are technically feasible, necessary, and worth the overhead in data storage and database administration. Given Internet access and Sybase front-end tools any collaborator can easily share our data in a tightly-controlled fashion that does not put an undue burden on the collaborator. Collaborators who wish to maintain their own Sybase database(s) can grant similar privileges to us, under their complete control. Work in progress with Johns Hopkins will allow GDB to access our physical mapping data using these methods.

(This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract No. W-7405-ENG-48).

S05
The Chromosome Information System

Suzanna Lewis, William Johnston, Victor Markowitz, John McCarthy, Frank Olken, and Manfred Zorn
Information and Computing Sciences and Engineering Divisions, Lawrence Berkeley Laboratory, Berkeley, CA 94720

Development of a comprehensive and general Chromosome Information System (CIS) is a major part of the computing effort at Lawrence Berkeley Laboratory's Human Genome Center. Molecular biologists interact with CIS through a graphical user interface to access and manipulate biological data without needing to know its underlying database structure and implementation. Currently CIS incorporates mapping information about chromosomes, maps (genetic, physical), markers (loci, probes), sequences (primers, STSs), and other relevant information (e.g., persons, citations, references to other databases). CIS has three main components, a graphical user interface, a database, and an intermediate layer of software that implements the translation from the data model of the user interface to the representation in the database. Separation of the user level, where actions on biological data are expressed, and the database level, where data are stored, enhances system flexibility and functionality.

The graphical user interface enables the user to manipulate various kinds of genomic information. Queries for retrieving maps can be formulated by selecting a region on the chromosome shown on the screen or by choosing from a list of available data. Graphical queries provide access to all data that pertain to a particular region on the chromosome for the novice user, whereas choosing from a list of entries gives experienced users rapid access to known information. The graphical presentation also allows users to visually compare different kinds of maps. Thus, genetic and physical maps of human chromosomes can be displayed side by side to reveal the different metric used with these kinds of maps. Information shown on the map can be followed either to reveal other relevant information, like contact persons or bibliographic citations, to trace data back to experiments that placed the object onto the map, or simply, to get more information about an object.

We have implemented a prototype user interface on a MacIntosh computer using SuperCard as the interface manager. Currently, we are working on porting the user interface to a Unix workstation running the X window system in order to obtain much better interactive response.

The database has been designed at a level of abstraction close to the biologists' view of this information. We have analyzed different types of biological objects, as well as their interrelationships, and organized them in a conceptual schema using the Extended Entity Relationship data model. The representation of maps has been chosen to accommodate any kind of map, i.e., genetic, physical maps. Any biological object represented in the database may have references, e.g, a literature citation or a pointer to another database, and persons, e.g., owners or contact persons, attached to it. The conceptual schema was described in ERDRAW, a specialized graphical editor, and then automatically translated by SDT into the lower level definitions that are used by the commercial relational database management systems.

Procedures that are stored in the database translate the biological objects into their respective representation in the database and link the database to the user interface.

The current implementation of the database uses a Sybase relational database management system on a Sun database server. The user interface is connected to the database server via a local area network.

For more information please contact Bill Johnston (wejohnston@lbl.gov) Bldg. 50B, Rm 2239, Lawrence Berkeley Laboratory, Berkeley, CA 94720.

For more information on ERDRAW and SDT refer to the poster "Tools for Creating and Manipulation Complex Objects" by V. Markowitz et al., January 16, 1991.

This work is supported by the Office of Health and Environmental Research Program and the Applied Mathematical Sciences Research Program, of the Office of Energy Research, U.S. Department of Energy, under Contract DE-AC03-76SF00098.

S06
Is Ethics Just the Bad and the Ugly?

Michael S. Yesley
Los Alamos National Laboratory, Los Alamos, NM 87545

An overview of (i) the ethical, legal and social implications ("ELSI") of the Human Genome Project, (ii) current ELSI activities in the United States and abroad, and (iii) DOE's program of research and education on ELSI. Discussion of why ethical inquiry may appear to focus on the harms and ignore the benefits that could result from the HGP.

Cloning/Amplification

S07

New Approaches for Constructing Expression Maps of Complex Genomes

R. P. Kandpal, S. Parimoo, S. Patanjali, J. Gruen, H. P. Arenstorf, H. Shukla, D. C. Ward, and S. M. Weissman
Yale University, New Haven, CT 06510

We have developed a procedure for preparing equal representation (normalized) libraries of cDNA fragments and used this to prepare such a library from human thymus. Preparations of such libraries from a number of other human sources including whole fetal brain and whole fetus are underway. These libraries are being used to identify the majority of coding sequences in large fragments of human DNA.

We have demonstrated that the combination of Not1 restriction endonuclease and BsuE methylase can generate very large fragments from human DNA, and that partial methylation can be used to generate mapping data around a Not1-BsuE site. We have further used this enzyme combination to isolate linking clones from an X chromosome specific genomic library. About 600 linking clones have been sequenced, representing about 120 different Not1 sites. About half of these have been regionally mapped on X by use of somatic cell hybrid lines.

Progress in combining the materials described above with previously described methods (1,2) to obtain expression maps for large DNA fragments will be reviewed.

[1]Kandpal, R.P., Ward, D.C. and S.M. Weissman. Selective enrichment of a large size genomic DNA fragment by affinity capture: an approach for genome mapping. (1990) Nuc. Acids Res. 18(7): 1789-1795.

[2]Kandpal, R.P., Shukla, H., Ward, D.C. and S.M. Weissman. (1990) Nuc. Acids Res. 18(10):3081.

S08

The Human X Chromosome: A Genome Strategy to Isolate Disease Genes by Positional Cloning

D. L. Nelson, M. F. Victoria, A. Ballabio, M. Pieretti, D. Kuhl, R. Bies, T. D. Webster, R. A. Gibbs, A. C. Chinault, S. T. Warren,** R. Nussbaum,*** and C. T. Caskey*
Institute for Molecular Genetics and *Howard Hughes Medical Institute, Baylor College of Medicine, Houston, TX 77030
**Division of Medical Genetics and Departments of Biochemistry & Pediatrics, Emory University School of Medicine, Atlanta GA 30304
***Department of Human Genetics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104

Over 400 yeast artificial chromosome (YAC) clones have been isolated from the human X chromosome, and ~150 of these have been assigned to regions defined by chromosome translocation and deletion breakpoints. Alu PCR (Nelson et al., Proc Natl Acad Sci USA 86:6686 (1989)) has been applied to YAC clones in order to generate probes, to identify overlapping clones, and to derive "fingerprints" and sequence data directly using total yeast DNA. Several clones have been described in regions of medical interest. One set of three overlapping clones has been found to cross a chromosomal translocation implicated in Lowe's syndrome. The identification of the gene involved in this oculocerebrorenal syndrome is ongoing. A second group of clones localized to deletion intervals close to the Fragile X mutation region. Analysis of one of these (RS46) has allowed isolation of the fragile site in an overlapping YAC clone and generation of a highly informative polymorphic marker assayable by PCR that exhibits no recombination with the Fragile X syndrome with an inferred LOD score of over 40. The region deleted in the contiguous gene syndromes of Xp22.3 includes the steroid sulfatase and Kallmann syndrome genes, and is another major focus. A candidate for the Kallmann syndrome gene is currently being analyzed. Other regions of interest on the X chromosome include X-linked lymphoproliferative disorder, Duchenne and Emery-Dreifuss muscular dystrophies, and the XIST locus in the X-inactivation region. The regional assignment of groups of YAC clones provides initiation points for further attempts to develop large cloned contiguous sequences, as well as material for investigation of regions involved in genetic diseases.

S09

A Bacteriophage T4 In Vitro DNA Packaging System To Clone Long DNA Molecules

Venigalla B. Rao, Vishakha Joshi, and Lindsay W. Black*
Department of Biology, The Catholic University of America, Washington, DC 20064,
(202) 319-5271
*Department of Biological Chemistry, University of Maryland Medical School, Baltimore,
MD 21201

Bacteriophage T4 packages about 170 Kb of its DNA that includes 2% terminal redundancy, in a strictly headful manner. We have recently purified the various packaging components of phage T4, and developed an _in vitro_ DNA packaging system. This system is being currently used to clone about 150 Kb size foreign DNA, and to construct genomic libraries. The 95 Kb _tryp_, and the 100 Kb _nad_ NotI fragments of _E. coli_ have been cloned into the P1-_lox_ vectors using this system (The P1-_loX_ vectors were developed by Dr. Nat Sternberg at DuPont). We have also generated genomic libraries of _E. coli_ DNA from partial _BamHI_ digests. The clones obtained were in the range of 40 Kb to 135 Kb. We are now addressing various aspects of this system, such as isolating large quantities of 150 Kb size intact DNA fragments, improving the efficiencies of cloning etc., in order to construct 150 Kb size human genomic libraries.

S10

Isolation of Human DNA Fragments Using a New Bacterial Cloning System

H. Shizuya, B. Birren, U.-J. Kim, V. Mancino, T. Slepak, and M. Simon
Division of Biology, California Institute of Technology, Pasadena, CA 91125

We have developed a new cloning system for the mapping and analysis of complex genomes. This system is based on the well studied prokaryote, _Escherichia coli_ and its plasmid, F-factor. Several lines of work on F-factors have suggested that they could be used for large mammalian DNA cloning. Large F′ DNA molecules in merodiploids have been shown to carry more than 1 mb of _E. coli_ DNA fragments. Moreover, the F-factor is stably maintained as a low copy number plasmid and its DNA replication is strictly controlled. We are developing the Bacterial Artificial Chromosome system to provide a supplement and an alternative to the Yeast Artificial Chromosome (YAC) system. The BAC vector (pBAC) consists of a 5.5 kb portion of the EcoR1 fragment F5 of F-factor, 1 kb containing the chloramphenicol resistant marker, the lambda _cos_ site, and synthetic

DNA containing the P1-loxP site, as well as HindIII and BamH1 cloning sites. Phage lambda terminase and P1cre enzymes cleave circular pBAC clones at cos and loxP sites respectively, and the linearized DNA may then be used for physical mapping after partial digestion with different restriction enzymes. The BAC host HS979 was specifically constructed for the stable maintenance of complex genomic DNA inserts in the pBAC plasmid by the removal of a) host-controlled restriction, b) methylation sensitive restriction, and c) various types of recombination systems including recA, recBC, and SbcBC. The system has being used to clone DNA from human cell lines as well as DNA from a hybrid cell line containing human chromosome 22.

S11

New Single Copy Amplifiable Cloning Vectors

Andrew A. Kumamoto and Philip Youderian
California Institute of Biological Research, La Jolla, CA 92037

We have constructed a set of vectors called STEALTH vectors that carry an S replicon, the Salmonella phage P22 early region, and a chloramphenicol-resistance determinant. The prototype STEALTH vector S110 can accept 250 kbp pieces of DNA as inserts, maintain them in single copy, and amplify them about 500-fold. We will discuss other features that we plan to add to these vectors to maintain the ends of human DNA inserts stably.

S12

Construction of Partial Digest Libraries from Flow-Sorted Human Chromosomes

L. L. Deaven, M. K. McCormick, C. E. Hildebrand, R. K. Moyzis, N. C. Brown, E. C. Campbell, M. L. Campbell, J. J. Fawcett, A. Martinez, L. J. Meincke, P. L. Schor, and J. L. Longmire
Center for Human Genome Studies and Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545

The National Laboratory Gene Library Project is a cooperative project between the Los Alamos and Lawrence Livermore National Laboratories. At Los Alamos, a set of complete digest libraries has been cloned into the EcoRI insertion site of Charon 21A. These libraries are available from the American Type Culture Collection, Rockville, MD. We are currently constructing sets of partial digest libraries in the cosmid vector, sCos1, and in the phage vector, Charon 40, for human chromosomes 4, 5, 6, 8, 10, 11, 13, 14, 15, 16, 17, 20, and X. Individual human chromosomes are sorted from rodent-human cell lines until approximately 1 μg of DNA has been accumulated. The sorted chromosomes are examined for purity by in situ hybridization. DNA is extracted, partially digested with Sau3A1, dephosphorylated, and cloned into sCos1 or Charon 40. Partial digest libraries have been constructed for chromosomes 4, 5, 6, 8, 11, 13, 16, 17, and X. Purity estimates from sorted chromosomes, flow karyotype analysis and plaque or colony hybridization indicate that most of these libraries are 90-95% pure. Additional cosmid library constructions and 5-10X arrays of libraries into microtiter plates are in progress. Libraries have also been constructed in M13 or blue-script vectors to generate STS markers for selection of chromosome specific inserts from a genomic YAC library. We have also been able to clone sorted DNA into YAC vectors and expected to be able to construct YAC libraries representing individual chromosomes.

Supported by the U.S. Department of Energy under contract W-7405-ENG-36.

S13

Tri- and Tetranucleotide Sequence Tandem Repeats-a PCR Mapping Tool

C. T. Caskey
Institute for Molecular Genetics, Baylor College of Medicine, Houston, TX 77030-3498

Abstract not available at press time.

S14

Closure of the Chromosome 19 Contig Map

Pieter J. de Jong, Chris Amemiya, Charalampos Aslanidis, Michelle Alegria, Jennifer Alleman, Chira Chen, Alex Copeland, Jeffrey Elliot, Emilio Garcia, Anne Olsen, and Anthony V. Carrano
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

As part of our effort to prepare a high-resolution physical map of human chromosome 19, we have analyzed 8813 cosmids representing a 5-fold chromosome redundancy by a fluorescent fingerprinting approach (abstract, Carrano *et al.*). Based on the fingerprints, the clones were assembled into 720 contigs containing 4922 clones (about 3 fold chromosome redundancy). The average contig contains about 7 cosmids and has 3.3 clones on the tiling path (about 85 kbp). The validity of contigs has been checked by a variety of approaches, including *in situ* mapping of multiple contig members to metaphase spreads (see abstract of Trask *et al.*) To expand the contig map, more clones are being fingerprinted. In parallel, we have initiated a more focussed approach towards mapping and linking cosmid contigs using chromosome 19-specific YAC clones and hybrid cells. To isolate YACs corresponding to known markers and contigs, we are using a copy of the arrayed library from Olson and coworkers (Washington University). To facilitate access of this library for screening with PCR primers (STSs) or for hybridization with *Alu*-PCR probes, the arrayed clones have been pooled by three independent pooling schemes (abstracts of Amemiya *et al.* and Alegria *et al.*). The array position of any YAC in the array is uniquely defined by a combination of pools derived from each of the different schemes. The use of the YAC pools for PCR screening eliminates the use of yeast colony hybridization as the last screening step and thus facilitates the screening of the YAC library. The pools are also used for hybridization screening. In this case, all pools are PCR amplified once with *Alu* primers. The PCR mixtures are then used as target DNA for Southern or dot-blot hybridization. As a result the 55,000 clone library can be completely represented by 96 superpools("Southern lanes") or 1728 pools ("hybridization dots"). The probes for the pool screening are also generated by *Alu*-PCR using isolated recombinants as templates. Rapid anchorage of the YACs to cosmid contigs is being achieved in a similar way. *Alu*-PCR probes are isolated from the YACs for subsequent hybridization to *Alu*-PCR products from cosmid pools (500 PCR mixtures spotted onto a small filter to represent 10,000 arrayed cosmids).The preliminary results support the usefulness of this approach (Amemiya *et al.*). *Alu*-PCR products are thus used in parallel to STSs as a "common language" to link cosmids, YACs and hybrid cells.

Work performed by the Lawrence Livermore National Laboratory under the auspices of the U.S. Department of Energy under contract No.W-7405-ENG-48.

S15

Physical Mapping of Human Chromosome 11

Glen A. Evans, David McElligott, Gary Hermanson, Licia Selleri, Susanne Maurer, Mary Saleh, Dan Kaufman, Kathy Lewis, Jun Zhao, Greg Huhn, Caryn Wagner, Shizhong Chen, Grai Andreason, Jim Eubanks, Maria Roman, Cindy Toraya, Ken Snider, Lisa Leonard, and Reece Hart
Molecular Genetics Laboratory and Center for Human Genome Research, The Salk Institute for Biological Studies, La Jolla, CA 92138

Chromosome 11 represents about 4.2% of the human genome and spans approximately 150 mb. We have derived a strategy for generating a physical map consisting of overlapping cosmid and yeast artificial chromosome clones utilizing six simple steps and based on organized cosmid and YAC libraries. Reagents utilized for this project include 1) a chromosome llql2-llqter cosmid library of 1000 members, constructed from a somatic cell hybrid in cosmid vector sCos-1, 2) a 18,000 member cosmid library constructed from a somatic cell hybrid representing the entire chromosome 11, 3) a total human genome YAC library and 4) a chromosome 11-specific YAC library prepared from a somatic cell hybrid. The steps in this analysis include: 1) regional mapping of cosmid landmarks by high resolution non-isotopic *in situ* hybridization to a precision of 1-2 mb, giving an average spacing of landmarks of 300-500 kb, 2) fine structure ordering of cosmids using *in situ* hybridization to interphase nuclei resulting in a resolution of less than 100 kb, 3) generation of DNA sequences from the ends of landmark cosmids by direct, double-stranded DNA sequencing from cosmid templates, 4) generation of PCR probes based on cosmid sequences, 5) isolation of corresponding yeast artificial chromosome clones (YACs) by PCR-based screening, 6) analysis of the resulting YAC clones for overlaps, sequence content and cosmid subcloning. In addition, cosmid libraries have been analyzed for a number of sequence structures including HTF islands, rare restriction sites, and polymorphic repetitive sequences, which may be placed in the emerging chromosome map. This analysis has already generated several combined cosmid/YAC contigs of greater than 2 mb. Further progress on this project will be discussed.

S16

Physical Mapping of Human Chromosome 16

N. A. Doggett, R. L. Stallings, M. K. McCormick, J. Dietz-Band, C. E. Hildebrand, L. L. Deaven, and R. K. Moyzis
Center for Human Genome Studies and Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545

Our goal is the construction of a complete physical map of human chromosome 16 (i.e., ordered 2 Mb contigs covering >95% of the 90 Mb euchromatic arms, with uniformly spaced STS markers; The U.S. Human Genome Project, The First Five Years FY 1991-1995, DOE/ER-0452P). Our strategy involves the rapid generation of DNA cosmid contigs representing approximately 60% of the target chromosome, followed by directed gap closure with yeast artificial chromosomes (YACs). The first phase of this goal, the rapid generation of "nucleation" contigs on chromosome 16, has been completed (Stallings, et al., Proc. Natl. Acad. Sci. USA 87, 6218-6222, 1990). Using an approach for identifying overlapping cosmid clones by exploiting the high density of repetitive sequences in human DNA, 553 contigs have been generated following the fingerprinting of

4,000 individual cosmid clones. This represents approximately 60% (54 Mb) of the euchromatic arms of chromosome 16, and was achieved with approximately one-fourth as many cosmid fingerprints as random strategies requiring 50% minimum overlap detection. By "nucleating" at specific regions, this approach allows a) the rapid generation of large (>100 kb) contigs in the early stages of contig mapping, and b) the production of a contig map with useful landmarks (i.e., $(GT)_n$ repeats) for rapid integration of the genetic and physical maps. All 4,000 fingerprinted cosmids have been rearrayed on high density filters. Such filters already provide investigators with access to greater than 90% of chromosome 16, with a 60% probability that any region is already present in a >100 kb contig. In collaboration with the laboratories of David Ward (Yale) and David Callen (Adelaide), eighty-five of these contigs have been regionally localized via in situ hybridization or somatic cell hybrid panels. The average "gap" size (containing only "singlets") is approximately 65 kb. Such gaps are easily closed with YACs. A single "walk" from each of the ends of our current contigs should, statistically, reduce the number of contigs to approximately 50, the 5-year goal of the Human Genome Project. To facilitate closure, we are constructing both a total genomic YAC library (from cell line GM130, using the vectors pJ597 and pJ598; currently 1X representation) and chromosome 16 YAC clones from monochromosomal hybrids and flow-sorted material. One hundred STS markers to key contigs are being generated. Current progress with YAC closure indicates that the complete physical map of chromosome 16 will be achieved in the next few years.

Supported by the U.S. Department of Energy under contract W-7405-ENG-36.

S17

Preliminary Correlation of the Physical and Linkage Maps of Human Chromosome 16

G. R. Sutherland, H. Kozman, D. F. Callen, H. Philips, T. Keith,* C. Julier,** J. C. Mulley
Department of Cytogenetics and Molecular Genetics, Adelaide Children's Hospital, North Adelaide, SA, Australia
*Collaborative Research Incorporated, Bedford, MA 01730
**Centre d'Etude du Polymorphisme Humaine, Paris, France

A physical map of polymorphic markers from chromosome 16 was constructed using a combination of Southern blot analysis, PCR analysis and *in situ* hybridization analysis of a panel of human-mouse hybrid call lines containing rearrangements of human chromosome 16.

A genetic linkage map consisting of 42 marker loci that had been physically mapped was also constructed. The data for each of the markers used in the construction of this map were obtained from Keith *et al.* (1987). Julier *et al.* (1990), Donis-Keller *et al.* (1987) and our laboratory (unpublished).

The order established from physical mapping provided the initial order for the genetic map. The map distances between adjacent loci were estimated using the computer program CILINK, from the LINKAGE package, assuming equal recombination fraction in males and females. Adjacent loci were inverted to obtain the odds against the alternative orders. If a marker was not able to be placed with odds of 100 to 1 or greater, the location score method, using the computer program CMAP, was used to determine if another position was more likely for a particular locus. In this manner, a multipoint genetic map was constructed spanning the entire length of chromosome 16. The total map length is 187.2 cM (sex average) and there is an average distance of 4.5 cM between marker loci. The order of the 42 marker loci was found to correspond on the physical and genetic maps.

S18

Progress Towards the Construction of an Interdigitated Physical and Genetic Map for Human Chromosome 19

K. M. Tynan, H. W. Mohrenweiser, A. S. Olsen, E. Branscomb, P. J. de Jong, and A. V. Carrano
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory,
Livermore, CA 94550

Utilizing a series of approaches, we have constructed a cosmid contig map of human chromosome 19 which spans an estimated 70% of the 60 million base pairs of this chromosome. Using a high resolution fluorescence-based, restriction fingerprinting method 8813 cosmids have been analysed; 4922 cosmids have been assembled into 720 contigs. Of these, 373 contigs have 4 or more cosmids present as minimum tiling path members; also included are 11 contigs which have in excess of 10 cosmids in the minimum spanning set. An important feature of the cosmid physical map is its interdigitation with known genetic information. We have currently mapped 41 of the over 100 unique markers assigned to chromosome 19 to cosmids. At present, 120 contigs have single genes, multigene families and repetitive elements assigned to them. Several of these gene/repetitive regions are now being studied in depth.

One particularly fruitful approach to assigning genetic markers to cosmids/contigs, is the use of conserved sequence motifs from multigene families. We have used a probe from the shared constant domain of the carcino-embryonic antigen (CEA)-like genes to probe our chromosome 19 enriched cosmid library. These genes, which map exclusively to 19q13.2, are members of a large family of 30–40 closely related genes, and are a subgroup of the immunoglobulin superfamily. Approximately 230 cosmids were identified as containing this conserved 'repeat unit' sequence. 197 of the CEA-positive cosmids have been assembled together with 103 additional cosmids into nine significant contigs which span approximately 1.2Mb (see abstract by Olsen et al.). We have also used a minisatellite sequence specific to the q13 region of chromosome 19, to identify approximately 440 cosmids. 240 of these cosmids have been assembled into 74 contigs. This minisatellite sequence has been found associated with 7 functional genes, ERCC1 ERCC2, PVS, XRCC1, ATP1A3, PRKCG and the previously identified APOCII gene.

We have completed the assignment of cosmids for eight of twelve markers necessary to construct a consensus physical framework map of chromosome 19. Four of the markers define loci which contain VNTRs while the remaining 8 define such genes as, INSR, LDLR, ATP1A3, CYP2A, APOC2, ERCC1, and PRKCG. STSs are being established for each of these loci.

Chromosomal translocations involving the short arm of chromosome 19 have been demonstrated in human hematolymphoid malignancies. In order to facilitate the generation of a high resolution physical map of this region, cosmids for ten oncogene and receptor markers assigned to the p-arm of chromosome 19, have been identified. These cosmids are being used to map the relative location of break-points in a series of leukemic cell lines.

The assignment of genetic markers to cosmids/contigs has and will continue to be an important feature of our evolving physical map for human chromosome 19. It allows us to define regions of interest as well as providing test cases for the assessment of contig integrity.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract No. W-7405-ENG48.

DNA Sequencing Instrumentation/Database Needs

S19

Large-Scale DNA Sequencing Needs

Leroy Hood, Robert Kaiser, Ben Koop, and Tim Hunkapiller
Division of Biology, California Institute of Technology, Pasadena, CA 91125

DNA sequencing is a multistep process, beginning with obtaining appropriate DNA clones and terminating with the detailed DNA sequence analysis of the information obtained. Each step in this multicycled process is a potential bottleneck for the overall sequencing process. Moreover, a data management system must be established to keep track of the various components in this complex process. We will discuss the major problems that arise in large-scale DNA sequencing and the particular computational needs that attend this process.

S20

High Resolution Restriction Fragment Separations by Capillary Electrophoresis

B. L. Karger, D. N. Heiger, M. Vilenchik, E. Szoko, and A. S. Cohen
Barnett Institute, Northeastern University, Boston, MA 02115

Capillary electrophoresis is emerging as a powerful tool for DNA separation and analysis. This instrumental approach to electrophoresis offers the possibility of high resolution, speed, quantitation, collection and automation. Our laboratory has developed columns utilizing sieving media within the capillary to achieve resolution of single stranded oligonucleotides and double stranded DNA.

In this paper we demonstrate the high resolving power of linear polyacrylamide columns to separate DNA restriction fragment digests in the range of 50 bp to several kbp. Low polyacrylamide concentrations are utilized, e.g. 3% monomer and 0% cross-linker, for separations. Such compositions cannot be used on slabs, since the anticonvective character of the sieving medium is too low. However, the walls of the capillary provide sufficient anticonvective properties for the successful use of such media.

Using a specially designed wall coating to minimize adsorption and electroendosmosis, we have achieved separation of restriction fragment mixtures in less than 10 minutes at 300 V/cm. In some examples, baseline separation of one or several base pair differences has been observed. When necessary, the polyacrylamide medium can be blown out of the capillary and reloaded. With this approach, the precision of relative mobility is less than 0.5% rsd. In this paper we shall explore a variety of separations and demonstrate the use of this method to rapidly determine restriction fragment length.

S21

Rapid DNA Sequencing Based Upon Single Molecule Detection

Lloyd M. Davis, Frederic R. Fairfield, Mark L. Hammond, Carol A. Harger, James H. Jett, Richard A. Keller, Babetta L. Marrone, John C. Martin, Harvey L. Nutter, E. Brooks Shera, Daniel J. Simpson, and Steven A. Soper
Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545

We are developing a laser-based technique for the rapid sequencing of 40-kb or larger fragments of DNA at rates up to several hundred bases per second. The approach relies on fluorescent labeling of the bases in a single fragment of DNA, attachment of this labeled DNA fragment to a support, movement of the supported DNA fragment into a flowing sample stream, and detection of individual fluorescently labeled bases as they are cleaved from the DNA fragment by an exonuclease. The ability to sequence large fragments of DNA will significantly reduce the amount of subcloning and the number of overlapping sequences required to assemble megabase segments of sequence information. Progress in each of the areas listed above will be discussed.

S22

Sharing Physical Map Data in the Genome Community

D. Nelson, T. Slezak,* R. E. Lucier,** P. L. Pearson,** J. W. Fickett, and E. W. Branscomb*
Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM 87545
*Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550
**Genome Data Base, The Johns Hopkins University, Baltimore, MD 21205

A preliminary means has been agreed upon by which map data generated at Los Alamos National Laboratory (LANL) and Lawrence Livermore National Laboratory (LLNL) may be presented to the community in the rich context of the Genome Data Base (GDB).

The national laboratories have now generated a significant amount of physical map data, and are beginning to integrate the physical with the genetic and cytogenetic maps. This map data will be of very general interest, especially as an increasing number of STSs are located on mapped clones.

Data sharing is of the essence of the genome initiative, yet a reasonable protection of proprietary interest is necessary, as is an acceptable economy of operation in the sharing process. It is the latter concern that we have begun to address.

The communication problem itself is made very much easier in this particular case by the fact that all three sites use the same relational database management system (RDBMS), namely Sybase; that Sybase already incorporates network communication between its user interface and data management components; and that all three sites are connected by the Internet. Differences in schema details, security issues, and data release policies all provide challenges, but these are being overcome.

In the first stage of this collaboration LANL and LLNL will place spanning sets of some of their more interesting contigs in conceptually separate collections on local machines, for export to GDB and the public. GDB will use Sybase to import this data at regular intervals. The collections for export will be maintained automatically from the primary collection. The latter will only be accessible locally.

This collaboration is a step towards making maps generated at LANL and LLNL available electronically to the genome community, in the appropriate context of other human genome mapping information.

S23

High Speed DNA Sequencing by Horizontal Ultrathin Gel Electrophoresis (HUGE)

Lloyd M. Smith, Robert L. Brumley, Eric Buxton, Howard Drossman, Anthony J. Kostichka, John A. Luckey, and David A. Mead
Department of Chemistry, University of Wisconsin, Madison, WI 53706

We have been exploring the utility of capillary electrophoresis for increasing the throughput of fluorescence-based automated DNA sequencing instruments for the past two years (1,2). Using this method it is possible to separate and detect fluorescently labeled products of DNA sequencing reactions up to 25 times more rapidly than in conventional electrophoresis. In order to extend this high speed separation to the parallel analysis of multiple samples we have recently developed an apparatus for performing electrophoresis in ultra thin (10-100 μm) slab gels. Fields as high as 300 V/cm can be applied to these gels, and readable sequence out to 320 bases is readily obtained from a 25 minute electrophoresis using autoradiographic methods. We have also designed and built a detection system for multiple fluorophore fluorescence-based sequencing in these ultra thin gels. This system uses a charge coupled device (CCD) detector operated in frame transfer mode for the real-time acquisition of fluorescence data. Preliminary data from this system will be shown.

1. Drossman, H., Luckey, J. A., Kostichka, A. J., D'Cunha, J. and Smith, L. M. 1990. High speed separations of DNA sequencing reactions by capillary electrophoresis. Analytical Chemistry, 62, 900-903.

2. Luckey, J. A., Drossman, H., Kostichka, A. J., Mead, D. A., D'Cunha, J., Norris, T. B. and Smith, L. M. 1990. High speed DNA sequencing by capillary electrophoresis. Nucleic Acids Research, 18(15), 4417-4421.

DNA Sequencing Methodology

S24

Efficient Sequencing of DNAs Cloned in λ Using Transposon Insertions and PCR Amplification

B. Rajendra Krishnan, Dangeruta Kersulyte, Claire M. Berg,* and Douglas E. Berg
Washington University, St. Louis, MO 63110
*University of Connecticut, Storrs, CT 06269

Bacterial transposons that insert quasi-randomly into target DNAs are valuable for genome analysis because they make mutations by insertion and serve as mobile binding sites for sequencing primers. We report here the use of direct and crossover (biparental) PCR to quickly map Tn5 *supF* insertions in phage λ clones, and thereby identify insertions from which to sequence cloned DNAs with minimal redundancy. The main steps in our protocol are:

1. Transposon insertion. Obtain insertions in lambda clones by growing phage on Tn5 *supF*-donor cells, and then selecting plaques on strain DK21, on which only *supF*-containing lambda can grow (see *PNAS* 86:5908-5912).

2. First-pass direct PCR mapping. Detect and map insertions in cloned DNA that are close enough to the left and right junctions with lambda to be PCR-amplified with vector-specific and Tn5 *supF*-specific primers. At least half of a typical cloned DNA target (≤ 21 kb) is covered in this first pass.

3. Crossover PCR mapping of insertions near the center of the cloned fragment. Map insertions in the remaining (central) part of the cloned DNA segment by crossover PCR between different insertion phages from the same mutagenized population: use a phage with the innermost known insertion in combination with other phages with unmapped insertions.

4. Linear DNA amplification sequencing. Treat PCR fragments with "GeneClean" to remove inhibitors, and mix aliquots with an end-labelled primer. Then use these DNAs as templates for multiple cycles of Sanger sequencing with Taq polymerase as in Murray, NAR 17:8889, 1989. Sequence ladders of more than 500 bp have been obtained.

In conclusion, the usefulness of bacterial transposons for analysis of cloned DNAs has been greatly improved by the development of PCR mapping and linear amplification sequencing of PCR fragments. The strategy for use with lambda phage clones presented here is efficient, extendable to larger cloned DNAs (in cosmid, P1 or F vectors), and amenable to multiplex methods and to automation. We anticipate that it will become a valuable alternative to traditional shotgun subcloning and primer walking methods for DNA sequence analysis.

S25

Transposon γδ (Tn1000)-Facilitated DNA Sequencing

Claire M. Berg
University of Connecticut, Storrs, CT 06269

Often it is easier to use transposons to deliver primer binding sites to large target DNAs than to subclone random DNA fragments next to a fixed site--the most frequently used method of template preparation. The transposon γδ has a number of properties that make it the most useful of the E. coli transposable elements for plasmid mutagenesis and sequencing since γδ:1) transposes efficiently, and generally randomly from one plasmid to another. Although γδ forms a cointegrate intermediate during intermolecular transposition, the cointegrate is broken down by a very efficient resolution system; 2) can be delivered to a target plasmid by conjugation from the E. coli F factor; 3) can be used to generate sequence information from both DNA strands; 4) has unique subterminal sequences; 5) transposes to new sites within a single plasmid (intramolecular transposition), with high efficiency and randomness, matching that of intermolecular transposition; and 6) transposes readily into DNA regions of high or low GC content (despite a preference for AT-rich sites); 7) does not require any gene in cis for transposition. tnpA and tnpR can be in trans (cloned in a plasmid or in wildtype γδ in the chromosome).

My collaborators and I (see posters presented by Wang et al., and by Berg and Strausbaugh) are exploring the use of mini-γδ derivatives for mapping and sequencing cloned DNAs and are developing an improved probe-based strategy for the rapid mapping of insertions of any transposon into plasmids.

S26
Computer Assisted Multiplex DNA Sequencing

G. M. Church, G. Gryan, S. Kieffer-Higgins, L. Mintz, P. Richterich, M. J. Rubenfield, and M. Temple
Howard Hughes Medical Institute, Department of Genetics, Harvard Medical School, Boston, MA 02109

In multiplex DNA sequencing (Science 240, 185), 40 sequencing reactions sets, each tagged with specific oligonucleotides, are pooled and run along with up to 60 other pools on a single gel and transferred to a membrane. 75 such membranes are hybridized simultaneously and reprobed 40 times. The resulting sequence film images are digitized by any of 5 scanner types. The computer program (REPLICA) uses internal standards from multiplexing to learn lane alignment and lane specific reaction rules. Automatic base assignments are then superimposed on the displayed images for editing. Images with overlapping DNA sequences are viewed side-by-side to facilitate discrepancy checking. Hash table guided dynamic programming routines for multi-sequence alignment of shotgun sequences in the megabase range are compatible in speed with the rest of the software (less than 1 hour per cosmid). Tests of these methods have involved over one megabase of primary multiplex data. Alkaline phosphatase-conjugated oligonucleotide probes provide non-radioactive visualization with hybridization and exposure times of less than 20 minutes. Direct transfer electrophoresis to nylon membranes has been refined, yielding readable sequences beyond 700 bases per lane-set with 24 lane-sets per gel and beyond 900 bases in one case using a 60 cm gel.

S27
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

Ligation-mediated genomic sequencing (LMGS) is a newly developed technique that allows high quality sequence information to be obtained from mammalian cells without cloning. In brief the technique is as follows. The first step is preparation of DNA with 5′ phosphorylated ends. Maxam-Gilbert cleavage provides such ends and generates sequence information that can be amplified by LMGS. Next, a gene-specific oligonucleotide (primer 1) is used for a primer extension reaction, generating molecules having a blunt end on one side. Linkers are ligated to the blunt ends, and then the linker-ligated molecules are amplified by an exponential PCR reaction done using the longer oligonucleotide of the linker (linker-primer) and a second, nested, gene-specific primer (primer 2). This method amplifies all molecules that have undergone complete primer extension and linker-ligation. Electrophoretic separation of the amplified fragments on DNA sequencing gels gives high quality sequence ladders which can be visualized by hybridization with an appropriate probe. The quality of the ladders obtained is the same for genomic mammalian DNA as it is for cloned DNA. As performed in our lab, the separated fragments are transferred to a nylon membrane prior to hybridization. This procedure has several advantages, including the possibility of a type of "multiplexing." LMGS will definitely be useful for sequences that are hard to clone, or where directional sequencing is desirable. The overall aim of this project is, as stated in the title, to improve and automate LMGS. Our first goals are to improve the chemistry and simplify the procedures; automation is a more long term goal.

S28

Sequencing with a Primer Library

F. William Studier and John J. Dunn
Biology Department, Brookhaven National Laboratory, Upton, NY 11973

Ability to sequence cosmid-sized or larger DNAs directly by primer walking, using only primers from a library, would improve efficiency and reduce sequencing costs by at least an order of magnitude over current practice. Sequencing could be done on a single preparation of DNA without subcloning, the sequence of both strands would emerge with little redundancy, and ambiguities could be checked easily by resequencing with a different primer. Since all the needed primers would be available in the library, such a system could provide the basis for developing very high capacity automated sequencing machines.

To have a manageable library, primers as short as nonamers or decamers must be able to prime sequencing reactions specifically and reliably, or some simple means of joining shorter primers to produce longer ones is needed. Work on factors affecting the ligation of short oligonucleotides will be presented.

S29

Combinatorial Mapping of Transposable Vectors for Sequencing Large Plasmid Inserts

R. Weiss and R. Gesteland
Howard Hughes Medical Institute and Department of Human Genetics, University of Utah, Salt Lake City, UT 84112

Efficient transposable vector systems for saturating plasmids with inserts containing common priming sites and multiplex identifier tags may have advantages over current popular vector systems, such as M13. These include: a robust *in vivo* method for producing clones which eliminates *in vitro* manipulations necessary for shotgun cloning of random fragments, bi-directional sequencing from a fixed point, and viable mapping strategies for isolating minimal spanning sets of inserts from random pools. Ordered inserts provide the minimal spanning set of clones required to complete a segment while correcting for potential hot-spotting of the transposon. Transposons also provide a method for delivering replication functions and selectable markers to large-clone inserts, which allows serial-sectioning of large plasmid inserts into smaller end-to-end clones. Utilization of these unique features of transposable vectors will reduce the effort required for going from a large-insert clone to finished DNA sequence.

Currently, initial sets of gamma-delta transposons are being used to sequence plasmids containing 10 kb inserts. The transposable vectors are conjuation-based systems that use co-integrate formation to select transposition events. Each transposon consists of 38 bp. terminal inverted repeats flanked by unique 16 bp. multiplex identifier sequences, common priming sites, an internal 34 bp loxP site required for resolution of co-integrates, and central rare-cutter restriction sites used in mapping. Two distinct configurations of these elements are being investigated: a minimal transposon containing these elements in a 260 bp. vector, and an expanded version carrying accessory functions (drug resistance and replication origins). These vectors are carried on an E. coli F plasmid pCJ105 and gamma-delta transposase is provided from a cloned gene in trans. Transpositions into plasmids are isolated by mating to an F recipient while selecting for plasmid and recipient specific drug resistances. Transposition frequency into cosmids of 4 x 10^4 transpositions/ml from a 3 hr. mating are obtainable.

Independent insertions are now being mapped and used as priming sites for sequencing 10 kb plasmids. These insertions are localized to small (100 bp) intervals by restriction site mapping of the insert site from combinatorial clone pools. The combinatorial approach allows mapping on a single gel the relative positions of 250 inserts within 10 kb fragments. The minimal spanning set is then processed through conventional Sanger dideoxy-sequencing in a multiplex fashion, allowing recovery of two divergent sequence ladders containing identifier sequence tags from each insertion.

New Techniques and Instruments

S30

Isolation of YAC Insert Ends by Homologous Recombination

Gary G. Hermanson, Merl Hoekstra,* and Glen A. Evans
Molecular Genetics Laboratory and *Molecular Biology and Virology Laboratory, Center for Human Genome Research, The Salk Institute for Biological Sciences, La Jolla, CA 92138

Yeast artificial chromosome clones provide a powerful technique for isolating and mapping large regions of human chromosomal DNA. Using probes derived from both extreme ends of a YAC clone insert, overlapping YAC clones can be isolated by "walking" resulting in the isolation of YAC contigs spanning several hundred kilobases. However, we have found that the most time consuming step in YAC walking is the reliable isolation of both ends of each YAC insert. To circumvent this problem, we have constructed two novel vectors that allow the rapid and efficient subcloning of the extreme ends of YAC clone inserts. These vectors function via homologous recombination in yeast followed by transformation and propagation in bacteria and are useful for any YAC clone isolated from any library constructed in pYAC derived vectors. The steps involved in YAC end subcloning are: 1) the transformation of the YAC-containing yeast culture with a vector that carries the LYS 2 selectable marker that allows homologous integration into one YAC arm, 2) selection for yeast cells that have homologously recombined the vector into the appropriate YAC arm by selection for lysine prototrophy, and 3) cleavage of the yeast genomic DNA with one of several restriction enzymes followed by recircularization, transformation, and antibiotic selection in bacteria. Once subcloned into bacteria, the YAC end can be directly sequenced and STS PCR primers synthesized, or the clone can be used as a hybridization probe for Southern blot analysis.

This system has several important advantages over other YAC screening methods. The end-clones isolated with these recombination vectors may contain up to several kilobases of DNA, sufficient for the synthesis of PCR primers which avoid highly repetitive regions. These vectors can use any of several available polylinker sites for the recircularization step and thus do not rely on the fortuitous placement of restriction sites or repetitive sequence elements in the YAC insert. The vapors incorporate a minimal amount of extraneous vector sequences upon recircularization so as to maximize the length of insert DNA that can be cloned. Finally, these vectors allow cloning of both ends of any YAC insert constructed in a pYAC derived vector in a very rapid and labor efficient manner.

S31

Double Minute Chromosomes as Multimegabase Cloning Vehicles

Peter J. Hahn, John Hozier,* and Michael J. Lane
State University of New York Health Science Center, Syracuse, NY 13210
*Florida Institute of Technology, Melbourne, FL 32901

A need exists for a cloning vehicle capable of maintaining DNA segments intermediate in size between whole chromosomes and YACs or cosmids. We have developed a system for using double minute chromosomes in mouse EMT-6 cells that can "clone" mammalian chromosomal fragments in the 1 to 10 megabase range suitable for sub-chromosomal library construction. To clone DNA in this size range, we first introduce linked Neo and dihydrofolate reductase (DHFR) genes into the genome to be sub-cloned. This is followed by lethal irradiation and fusion to mouse EMT-6 cells, and selection for neo resistance. Subsequent selection for increasing levels of methotrexate (MTX) resistance leads to amplification of the introduced segment. To demonstrate the feasibility of this system, we have co-transfected Neo and DHFR genes into Chinese hamster ovary (CHO) cells, and transferred the CHO fragment containing the linked genes to mouse EMT-6 cells by radiation/fusion hybridization, and amplified the introduced chromosomal segment by selection for resistance to increasing levels of MTX. We have used pulsed-field gel electrophoresis to map approximately 700 kb surrounding the introduced Neo, DHFR genes in the CHO donor line and in five EMT-6 radiation fusion isolates that received the same CHO chromosomal segment. Four of the five radiation/fusion hybrids received segments with maps indistinguishable form the donor - the fifth presumably had a radiation breakage within the 700 kb segment we have mapped. No further rearrangements have been detected in the subsequent amplification steps. All lines contain cytogenetically observable double minute chromosomes, and the introduced genes are unstable in the absence of the selection. These data suggest that this system should be ideal for radiation/reduction type sub-chromosomal libraries because chromosomal fragments not attached to the selective marker will be readily lost with time, and the small size of double minute chromosomes should facilitate isolation of the introduced DNA.

This genome fragmentation strategy should allow construction of comprehensive human genome libraries containing less than 10,000 members while retaining high redundancy. We report the methodology we are employing to construct a partial library of human chromosome 17q. The process involves infection of a somatic cell hybrid containing human 17q with the defective retrovirus pZipneo, lethal irradiation and screening of several hundred G418 resistant EMT-6 recipient isolates for the presence of human DNA. Characterization of several of the human DNA containing EMT-6 cells identified, using inter-Alu PCR and *in situ* hybridization, reveals that this strategy can be used to produce selectable human double minute chromosomes in the mouse EMT-6 cell line.

S32

Labeling DNA with Stable Isotopes: Their Potential in Sequencing DNA

K. Bruce Jacobson
Biology Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831

To apply stable isotopes in DNA sequencing there are three topics that must be addressed: 1) labeling the DNA, 2) separating the sequencing fragments, and 3) detecting the separated, labeled DNA fragments. Resonance ionization spectroscopy (RIS) using a mass spectrometer has been used routinely to detect individual isotopes in complex matrices. Since iron and tin isotopes are obtained as the oxides, chemistry has been developed to convert Fe_2O_3 to ferrocene carboxylic acid and SnO_2 to (triethylstannyl)alkanecarboxylic acid. These acids were converted to the N-hydroxysuccinimide esters and allowed to react with oligonucleotides that were synthesized so as to contain a hexylamine on the 5′-terminus. Tin derivatives bearing two NHS ester groups have also been obtained and will be useful for double-labeling experiments. Both iron and tin-labeled oligonucleotides were electrophoresed on polyacrylamide gel. Two modes of using RIS were compared for detecting these elements, either on the dried polyacrylamide gel directly or after transferring the DNA to a Nylon membrane. Sputter initiated RIS (SIRIS) and laser atomization

RIS (LARIS) both proved successful in detecting iron-labeled and tin-labeled DNA on polyacrylamide but only the latter was successful with Nylon. Specific examples of separation of 24-, 25-, and 26-mers and the detection of the ^{116}Sn and ^{118}Sn by which they were labeled will be presented. Faster sequencing procedures will be available since 1) multiplexing of the electrophoresis process occurs when many isotopes are used simultaneously and 2) the RIS can be operated at 6 kHz and should be able to analyze 500 DNA electrophoresis bands in less than 3 seconds. (Research sponsored by OHER, U. S. DOE under contract DE-AC-05-84OR21400 with Martin Marietta Energy Systems and DE-AC-05-89ER80735 with Atom Sciences, Inc.).

Abstract for DOE Human Genome Program Workshop, February 17-20, 1991; Santa Fe, NM.

S33

Two-Color Fluorescence *In Situ* Hybridization Mapping of Chromosome 19

B. Trask, A. Fertitta, M. Christensen, B. Brandriff, L. Gordon, A. Copeland, K. Tynan, and A. V. Carrano
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

Fluorescence in situ hybridization (FISH) has three main roles in the construction of a long-range map of chromosome 19: (1) localizing cosmids to chromosomal bands, (2) checking contig construction, and (3) ordering contigs. Cosmids are non-radioactively labeled with biotin or digoxigenin. Probe hybridization sites are labeled with fluorescent tags (e.g. Texas Red or FITC (green), which can be viewed simultaneously. Using a combination of interphase and metaphase targets, sequences separated by 50 kbp and greater can be ordered. Hybridization to pronuclei, more highly decondensed targets, is described in a separate abstract (Brandriff et al.). (1) As of 1/3/91, we have regionally localized ~302 cosmids to cytogenetic bands on chromosome 19. 82 of these are associated with genes or genetic markers. 242 fall in 105 contigs, which are well distributed along chromosome 19. (2) The validity of our contig construction by fingerprinting and overlap detection algorithms is assayed using FISH. For each contig tested, two or more contig members, often those at opposite ends of the contig, are hybridized to metaphase and interphase chromatin and labeled in different colors. False contigs are identified as those whose members map to different bands or > 1 μm apart in interphase chromatin. Of 65 contigs checked by FISH, we have identified 2 (3%) that contain a false join. This percentage is expected to decrease as more cosmids are entered into the contig solution. (3) Contigs that map within the same band are ordered by measuring their proximity in interphase chromatin. Interphase distance increases with genomic distance. Sequences separated by >100 kbp and < 2 Mbp are most easily ordered with this approach. To date, the order of 10 contigs mapped to 19q13.2 has been determined by FISH mapping.

Work performed under the auspices of the U.S. D.O.E. by LLNL under contract # W-7405-ENG-48.

S34

Development of a Human Viral-based Genomic Library of 150–200 Inserts Size

Tian-Qiang Sun and Jean-Michel H. Vos
Department of Biochemistry and Biophysics and Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27514

A human viral-based library of genomic DNA of large size inserts would be a very useful resource to map, isolate, sequence and assay genes and other functional genomic regions. Progress will be presented on the development of an infectious Epstein-Barr virus-based (EBV) human genomic library of 150-200 kb size inserts which is propagated exclusively in human cells. Using a mini-EBV as genomic cloning vector and a resident EBV from human lymphoblastoid cells (HLC) as helper virus, we have shown by Pulse Field Gel Electrophoresis analysis that between 175 and 225 kb of engineered DNA can be packaged into infectious EBV virions. We observed by Fluorescence Activated Cell Sorter analysis that EBV virions carrying the bacterial betagalactosidase gene can efficiently infect HLC. The mini-EBV DNA inserts which is established as extrachromosomal plasmid in HLC can be recovered as infectious EBV; as a test case, we have been able to shuttle mini-EBV cloned DNA into HOC from several Fanconi's anemia patients, a syndrome characterized by high genetic instability. We propose that the use of such EBV-based library in combination with the appropriate genetic complementation strategy should allow the isolation and/or assay of genes from various human syndromes, including Fanconi's anemia.

Poster Abstracts

Informatics: New Methods, Unresolved Problems

Cloning/Amplification

Cloning/Amplification (continued)

Mapping: Progress and New Methods

DNA Sequencing/Instrumentation Database Needs

DNA Sequencing/Instrumentation Database Needs (continued)

DNA Sequencing Methodology

New Techniques and Instruments

New Techniques and Instruments (continued)

Poster Abstracts

Informatics: New Methods, Unresolved Problems

P01

Making the LLNL Genome Database 'Biologist Friendly'

Linda K. Ashworth, Tom Slezak, Mimi Yeh, Elbert Branscomb, and Anthony V. Carrano
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

Large amounts of data generated by the LLNL Genome Project are stored on Sybase, a relational database. To date, seventy-five tables contain some 100 Mb of data. These tables include descriptive information about libraries, clones, genetic loci, and experimental design which are linked to tables containing results from fingerprinting, hybridizations, PCR, and restriction mapping.

In addition to the data input and retrieval tools provided by Sybase, we have designed and are continually updating tools that make data easy to input, access, and use by biologists who may have minimal contact with formal database theory.

These tools include: a published schema, (flow charts of the database design, showing tablenames and fieldnames within tables); a data dictionary describing each field and its relation to others; "dataview" tables that allow users to view data stored in multiple tables simultaneously; and scripts containing complex database queries which are requested by the biologist. For example, using a script, it is simple to query the database for all cosmid clones found in contigs which *in situ* hybridize to any probe for a given locus.

Creation of new tables, dataviews and scripts is an interactive process between biologists and computations staff. This insures biologist design input, and results in tools that serve specific well defined needs. In the process, some users are becoming familiar with SQL language, allowing them freedom to write simple queries independently.

The Human Genome Browser (see abstract by Wagner and Slezak) has also been built to view physical map data graphically. This interactive tool is user friendly, and is now used by all biologists involved with physical and genetic mapping at LLNL.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory, Contract No. W-7405-Eng-48.

P02

Computation and Analysis in Support of Mapping Complex Genomes

David Balding, Tom Blackwell,[1] Randall Dougherty,[2] Frederic Fairfield,[3] Jim Fickett,[3] Karen Schenk,[3,4] David Torney,[3] and Clive Whittaker[3]
University of London, Queen Mary and Westfield College, London, United Kingdom
[1]Department of Statistics, Harvard University, Cambridge, MA 02138
[2]Department of Mathematics, Ohio State University, Colombus, OH 43210
[3]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545
[4]University of California, Berkeley, CA 94720

A topic of interest was the analysis of fingerprinted clones. We derived formulas for the likelihood of DNA fingerprint data with nondetected fragments. The length dependence of fragment size measurement reproducibility and the reproducibility of Southern Blot hybridization of repetitive sequence probes to restriction fragments of cosmid clones from Human Chromosome 16 were derived from the data and used in the detection of clone overlap. Algorithmic and structural changes in our computer program developed to determine clone overlap probabilities from fingerprint data enable this program to be used for a variety of fingerprint data.

Heterogeneity present in GenBank® DNA sequences was analyzed. Modeling was used to determine the effects of sequence heterogeneity on the contig size distribution for a variety of mapping strategies; the computer programs used to carry out this modeling were documented and widely disseminated.

A data structure developed by Tarjan for determining the connected components of a graph at all thresholds was modified to determine likely clone order within contigs and to reject false positive overlaps.

Finally, theoretical predictions were derived for STS mapping progress as a function of the number of STSs and clones used.

P03

Support for Data Management and Data Sharing in Genome Mapping Projects: The Tin Standard

E. W. Branscomb, T. R. Slezak, and A. V. Carrano
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

Efforts to construct detailed clonal physical genome maps generate relatively large assemblages of data, a large fraction of which (aggregating to perhaps 1-10 Gbyte of storage per 100 Mbase mapped to 40Kb resolution) must be saved and made usefully accessible. The value returned from the investment of public funds in such efforts is strongly influenced by the way in which the associated data storage and data management functions are carried out.

The main issues that determine value and also depend strongly on the design choices are: (1) the ability to provide for good QC, reliable sample and reagent tracking and general project management; (2) the permanence, completeness, and safety of the data storage; (3) the ability to provide adequately powerful and user-compassionate facilities for data entry, data recovery and general data query; (4) the ability to provide to the scientific community and to external collaborators, facile and timely access to critical results and supporting data; (5) the ability reliably and easily to protect data against undesired access and modification; (6) the ability to integrate data and the data base (DB) itself with other related, but remote, DBs; and (7) the cost of meeting these needs. Notwithstanding the importance of these issues, virtually all of the decisions involved are at present highly controversial.

The design questions relating to these performance issues will be discussed, using as a straw man the particular choices we have made at Livermore. These design questions include:

- The DB model used (e.g. relational vs object oriented); the wisdom/feasibility of storing physical mapping data in relational models, and of using "research" RDBMS or incompletely "developed" commercial systems.

- The value of using a DBMS designed on a network-compatible "client-server" architecture.

- The need for "real" computers (Unix, virtual memory, memories expandable to 30 to 60 Mbytes, fast Gbyte disk systems, 20-50Mips) vs "personal" computers --- as DBMS servers.

- The need for collections of computers connected on ethernet LANS and having "direct" INTERNET connectivity.

- The value/danger in using the currently popular "majority choice" commercial RDBMS (i.e. Sybase).

- The size of the hardware and manpower investment needed.

- The control of data access and data security.

- The value of making "distributed joins" and achieving "transparent" "integrated" remote data access.

The hardware and software components of an example "TIN standard" DB system will be described and the meaning, intent and implications of its proposed use as a "standard" will be discussed.

Work performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. W-7405-ENG-48.

P04

SCORE: A Program for Computer-Assisted Scoring of Southern Blots

T. M. Cannon, R. J. Koskela, C. Burks, R. L. Stallings, A. A. Ford, and J. W. Fickett
Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM 87545

To speed up Southern blot scoring in the construction of a chromosome 16 physical map (see abstract by Moyzis *et al.*), we developed SCORE, a program to automate many routine steps formerly done by hand (sliding films over one another, counting bands and lanes, typing the results into a database). The LANL fingerprinting technique is based on two single and one double digest of each clone, and uses both the fragment lengths and certain repetitive element probe hybridizations to partially characterize the clone. After the fragment lengths are stored in the local Laboratory Notebook database (see abstract by Nelson *et al.*), SCORE is used as follows.

SCORE retrieves the fragment lengths from the Laboratory Notebook and constructs a synthetic image of the ethidium bromide stained gel from which these fragment lengths were determined. The user then reads in the autoradiogram image from the Southern blot on a desktop scanner, and "stretches" it (by a simple affine transformation) to precisely overlay the synthetic image already on the console screen. Then for each fragment band SCORE offers the user a menu of possible signal strengths, from which one is chosen. This choice remains on the screen, and may be revised. If a band shows on the autoradiogram which does not correspond to any fragment in the database, a new fragment may be added. When all bands have been scored to the user's satisfaction, SCORE will store all the hybridization signals (and any new fragments) directly into the Laboratory Notebook.

This program has speeded the scoring of the Southern blot autoradiograms by about a factor of seven, and has also resulted in greater accuracy and an overall reduction in the proportion of tedious tasks. SCORE is written in C, Fortran, and SQL, and requires a Sun workstation with a color monitor.

P05

Integration of Automated Sequence Analysis into Mapping and Sequencing Projects

C. A. Fields, C. A. Soderlund, and P. Shanmugam
Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003-0001

The goals of the **gm** automated sequence analysis project include both the development of tools to make sequence analysis faster and more informative, and the discovery of new features of sequences that can be used to facilitate the analysis of large DNAs that may contain many genes. During the last year we have added a number of functions to **gm**, developed several new interactive analysis tools, and used the system to analyze sequences, primarily from the nematode *C. elegans,* obtained from GenBank® or from collaborators at other laboratories. Some of the results of this work are as follows.

- Functions for displaying cDNA data, restriction maps, and STSs have been added to the **gm** interface. These allow **gm** to be used for a variety of experimental design tasks. **gm** can also now predict structures of genes given partial cDNA data (see accompanying abstract by Soderlund et al.).

- **gm** has been used extensively in the analysis of the 54 kb *unc-22* cosmid from *C. elegans,* which was sequences by Guy Benian and colleagues at Emory University. This cosmid contains at least 5 different genes; two of these were predicted by **gm** independently of their experimental discovery. Working with this cosmid has been very valuable for understanding the behavior of sequence analysis methods on sequences containing multiple genes.

- The **gm** graphic interface is being used for the display and manipulation of exon maps in the *C. elegans* Community System, a distributed database of genomic information on *Caenorhabditis* being developed by B.Schatz and S. Ward at the University of Arizona. The interface displays exon maps, restriction maps, the DNA sequence, and predicted protein sequences in a single window; it thus provides access to both physical and functional information at high resolution in the Community System.

- A variety of new tools for calculating information contents of sets of aligned sequences have been developed. These are being used to examine the information contents of both single-base positions and base correlations in splice sites of *C. elegans* and *Drosophila.* Significant differences exist in the information contents of introns from short versus long introns in both organisms.

- New tools for sensitive base-composition analysis of sequence regions have also been developed. These are currently being used to better characterize both exon and intron sequences in *C. elegans* and humans.

As additional tools are developed and tested, they will be distributed to the community as part of the **gm** software package.

P06

Efficient Algorithms for Multiple Sequence Alignment with Guaranteed Error Bounds

D. Gusfield
Computer Science Division, University of California, Davis, CA 95616

Multiple string (sequence) alignment is a difficult problem of great value in computational biology. There are two rather different ways to measure the goodness of a multiple alignment, depending on whether one wants to find conserved subpatterns in the set of strings, or to build a tree deriving the strings from a common ancestor. For both measures, no efficient algorithm is known for optimally solving the problem for any but very small cases. A common approach in computer science to computationally hard problems is to develop fast (heuristic) algorithms whose maximum possible deviation from the optimal solution can be *proven* to be bounded by a small multiplicative factor. For the multiple string alignment problem (under either measure), no bounded error methods have been reported. In this work we provided the first such bounds. We developed two computationally efficient multiple alignment methods (one for each objective function) whose value is guaranteed never to be more than twice the value of the optimal multiple alignment. For both objective functions, the guaranteed bounds are even smaller when the number of strings is small (1.33 for three strings of any length), and for one of the objective functions, the method also yields a non-obvious *lower bound* on the value of the optimal solution. In addition to their direct use in producing multiple alignments, these methods and their associated bounds, have several indirect uses as well.

Supported by DOE grant DE-FG03-90ER60999 July 1990–January 1991.

P07

A New Algorithm for Constructing Suffix Arrays

D. Gusfield
Computer Science Division, University of California, Davis, CA 95616

The suffix array and its use in pattern searching was recently introduced by Manber and Myers as a more space efficient alternative to a suffix tree. For large alphabets, the suffix array is a valuable contribution, with natural applications in computational biology, where many important problems (such as pattern matching of restriction enzyme maps) have large "alphabets." In this work we give a different algorithm for constructing a suffix array and its associated *Lcp* (longest common prefix) values. Our algorithm is as fast ($O(n \log n)$) as the previous algorithm, but reverses the conceptual burden. In the Manber-Myers approach, the algorithm for building the suffix array is conceptually simple, while the one for finding the *Lcp* values is complex; in our approach, the suffix array construction is more involved, while computing the *Lcp* values is done simply in $O(n)$ time. Further, our approach provides much more information about the suffixes of the string, and hence can be easily extended to compute a full suffix tree, rather than just a suffix array.

Supported by DOE grant DE-FG03-90ER60999 July 1990–January 1991.

P08

PARAL: A Software Package to Efficiently and Optimally Align Sequences Using Parameterized Match, Mismatch, Indel and Gap Weights

D. Gusfield, K. Balasubramanian, D. Mayfield, and D. Naor
Computer Science Division, University of California, Davis, CA 95616

In the central problem of aligning two sequences (strings) there is considerable disagreement about how to weigh (score) matches, mismatches, insertions, deletions, and gaps. PARAL allows the user to avoid specifying exact weights by (provably efficiently) computing optimal alignments as a *function* of *variable* weights. The central use for PARAL is to study the sensitivity of the problem to choices made for the weights, and to give a small set of alignments such that for any choice of α, β, one of the alignments is optimal, and can be found quickly. PARAL presently runs on a Sun Sparc station. For example, consider the problem of aligning two strings to maximize (# of matches – α # of mismatches – β # indels), where α and β are variables. In one mode, the user inputs a choice for α and β, PARAL computes an optimal alignment A for that choice, and then determines and displays the region P in α, β space such that A is an optimal alignment for any point in P, and for *no* points outside of P. That space can be proved to be a *convex polygon*. In a more complex mode, PARAL completely partitions the α, β space into such convex polygons.

Supported by DOE grant DE-FG03-90ER60999 July 1990–January 1991.

P09

Mathematical Results About the Partition Computed by PARAL

D. Gusfield, K. Balasubramanian, D. Mayfield, and D. Naor
Computer Science Division, University of California, Davis, CA 95616

In addition to the program to find the polygon partition, we partially characterized the partition structure. There is a crucial distinction between the case when all spaces are counted in the alignment, including terminal spaces, and the case when they are not. In the case that all spaces are counted, we proved that all the polygons are infinite; that the number of them is at most $n^{2/3}$; that each line that bounds a polygon must be of the form $\beta = c + (c + 0.5)\alpha$ for some c; and that the entire decomposition can be found in $O(knm)$ time, where k is the actual number of polygons and n and m are the lengths of the strings. Hence each polygon can be found nearly as quickly as the best algorithms find a *single* alignment between two strings (i.e. with a constant factor). We have similar results for more complex objective functions, for example when gap penalties are added (a gap is a consecutive interval of spaces). We also considered the problem when the weight of a mismatch is given as a times a particular constant that depends on the pair of characters (as in the Dayhoff matrix). In this more complex case we proved that along any line in the α, β space, the number of polygons is sub-exponential. We also developed an algorithm for this more complex case which finds the decomposition in $O(nm^2 \log n)$ time per polygon.

Supported by DOE grant DE-FG03-90ER60999 July 1990–January 1991.

P10

Image Acquisition, Management, and Processing at the LBL Human Genome Center*

William Johnston, Antony Courtney, Marge Hutchinson, Joe Jaklevic, Bill Kolbe,
David Robertson, Brian Tierney, and Ed Theil
Information and Computing Sciences and Engineering Divisions, Lawrence Berkeley Laboratory,
Berkeley, CA 94720

We have developed an integrated imaging system for the LBL Human Genome Center in order to support production chromosome mapping. Many of the basic hardware and software tools for image acquisition, indexing, storage, and analysis have been adapted from existing systems and integrated with locally produced software.

In the current generation of laboratory experiments images typically arise from film based auto-radiogaphy. Incorporation of this data into computer information systems involves converting sufficient information from film image to digital image, and characterization of each image in terms of relevant experiment parameters. Image capture is routinely done with medium to high quality CCD camera-based digitizing systems.

The next step in the evolution of image data is to eliminate the film step and directly image the experiment. Technologies for both auto-radiography imaging and light emission based imaging (both fluorescence and luminescence) now exist (with the advent of phosphor image plates, and sensitive, low noise CCD cameras) and are in use at LBL. Use of these direct, volatile imaging methods implies that indexing and storage on reliable media must be an integral part of any production environment that relies on electronic imaging. LBL is addressing these issues through use of image data management systems and archival mass storage systems.

Definition of the image database, and entry of data into that database is handled by a front-end system. The image database and query system are general, but the structure of information in a particular database is specific to the type of experiment that gives rise to that set of images. The front end data entry system assists in entering accurate and complete information into the database.

Biologists access images in the database by forming queries using pull-down menus that list the experimental parameters. Retrieved images can be inspected visually and used as input to automatic feature detection software and quantitative analysis software. The mechanism of identifying sets of images according to some criteria is kept deliberately separate from subsequent operations on the images. Programs that display and analyze the images neither know nor care how a particular image was selected. This separation of function is akin to being able to use standardized parts in the construction of a complex object, and greatly simplifies the design, and enhances the flexibility of the system.

The volume of data, dispersed user community, and mass storage systems remote to biology laboratories also implies that imaging devices, permanent storage mechanisms, analysis workstations, and the software that ties all of these together must all run in a network based computing environment. LBL's lab wide network (including TCP/IP, NFS and ethernet) is used for this purpose.

*This work is supported by the Office of Health and Environmental Research Program and the Applied Mathematical Sciences Research Program, of the Office of Energy Research, U.S. Department of Energy, under Contract DE-AC03-76SF00098.

P11

Statistics of Sequence Matching

E. L. Lawler and W. I. Chang
Computer Science Division, University of California, Berkeley, CA 94720

The minimum number of substitutions, insertions, deletions needed to transform one sequence into another is called *edit distance*. It is a surprising fact that relatively little is known about the average case behavior of edit distance. While early works of V. Chvátal, D. Sankoff, and J. Deken on *longest common subsequence* implies the expected edit distance between two uniformly random sequences of length m (as $m \to \infty$) is $C_b m$ for some constant C_b that depends only on b the alphabet size, there has been no "formula" given in the literature for C_b. Since any match to much fewer than C_b-fraction differences can be considered "significant," a basic understanding of these constants is of paramount importance. The difficulty lies in the fact that the algorithmic formulation of edit distance (like approximate matching) is highly recursive; the natural Markov model has exponentially many states. We have found a way to subdivide the Markov model. Although this is not fully proved, we strongly believe C_b is about $(\sqrt{b} - 1)/\sqrt{b}$. Furthermore, we conjecture that the expected minimum value of row m of the dynamic programming table for approximate matching is $(1 + o(1))C_b(m - \log_b n)$. If proved, such a result leads to linear expected time, polynomial space (in the length of the pattern) algorithms allowing a *high constant fraction* of differences in a match.

Supported by DOE gant DE-FG03-90ER60999
July 1990–January 1991.

P12

Tools for Defining and Manipulating Objects in Biological Applications[1]

Victor M. Markowitz, Arie Shoshani, and Ernest Szeto
Information and Computing Sciences Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720

As part of an effort to facilitate the management of data in biological applications, we have developed tools and methodologies that support the definition and manipulation of data in terms of *objects*. The definition and manipulation of objects involves highlevel abstract constructs that are natural to biologist users. The data on these objects is stored using commercial relational database management systems (RDBMS). The tools support the definition and manipulation of objects, and translate them into RDBMS definitions and manipulations. We briefly describe below the object definition and manipulation tools.

SDT is an object-oriented tool for developing database applications on RDBMSs. The purpose of SDT is to provide a powerful and easy to use object definition interface for users who are not data management experts, by insulating the schema designers from the underlying RDBMS. Using SDT increases the productivity of the database design process, and simplifies subsequent schema changes.

The schema design tool uses the *Extended Entity-Relationship* (EER) data modeling methodology. The EER model includes in addition to the basic construct of object, both generalization and aggregation abstraction capabilities. EER schemas can be specified and edited via a textual language specification, or by using a graphical schema editor called ERDRAW. Once an EER schema is specified, SDT is employed to generate schema definitions for specific RDBMSs, including

SYBASE 4.0 and INGRES 6.3. These definitions include specification of referential integrity constraints, using *triggers* in SYBASE and *rules* in INGRES.

SDT allows the specification of both structural and descriptive information regarding schemas. Thus, the information about schemas is stored in a *metadatabase*, using the underlying RDBMS. SDT generates both the schema definition and the data manipulations necessary to load this metadatabase.

We are currently developing a tool called QUEST for assisting users in interactively specifying database queries in terms of objects. QUEST is designed for users who are not familiar with database management systems, query languages such as SQL, and operators such as *join*. Instead, users are only aware of the existence of objects and attributes, and are guided in the process of specifying queries.

This process consists of selecting items (objects, attributes, operators, and values) that are presented on a workstation screen. QUEST will allow both textual and graphical specifications of queries.

Like SDT, QUEST is designed to be employed on top of commercial RDBMSs, and to take advantage of commercially available software. Accordingly, queries specified using QUEST are subsequently translated into SQL queries. QUEST uses the metadatabase generated by SDT for inferring the connections (paths) between the objects specified in object-oriented queries, and for mapping information to translate these queries into SQL queries.

SDT and ERDRAW have been applied to the design of several databases at LBL, including the Chromosome Information System (CIS) database.

SDT and ERDRAW are implemented on Sun workstations under Sun Unix OS 4.1, using C, LEX, and YACC for SDT, and C, and the X11, XView toolkit for ERDRAW. SDT and ERDRAW are described in reference manuals published as technical reports LBL-27843 and LBL-PUB-3084, respectively.

QUEST is currently being developed using C, LEX, and YACC. The graphical query interface is being developed using C++ and the InterViews and Unidraw class libraries.

[1] This work is supported by the Office of Health and Environmental Research Program and the Applied Mathematical Sciences Research Program of the Office of Energy Research, U.S. Department of Energy, under Contract DE-AC03-76SF00098.

P13

On the Number of Solutions to the Probed Partial Digestion Problem

D. Naor
Computer Science Division, University of California, Davis, CA 95616

The *Probed Partial Digestion* (PPD) method partially digests the DNA with a restriction enzyme (RE). A probe, located between two RE cutting sites, is hybridized to the partially digested DNA, and the sizes of fragments which the probe hybridizes to are measured. The objective is to reconstruct the linear order of the RE cutting sites from the set of measured lengths. Since the reconstruction is not always unique, we are interested in the question of how many distinct solutions there are to a PPD reconstruction instance. We study this problem under the assumption that the data are complete and accurate (i.e. fragment sizes are integers).

We first observe that this is equivalent to asking a related question: given a set S of integers, how many distinct pairs of sets (X,Y) can there be so that $S = \cup_{i,j}\{y_j + x_i\}$. We then show that, for infinitely many numbers N, there is an input set S to the PPD problem of size N which has $0.5N/\sqrt{0.5\pi \log N}$ distinct feasible solutions if X and Y must be of the same size, and $0.5N$ feasible solutions for any X and Y. These bounds hold even in the restricted case in which a number cannot appear more than once in S, X and Y. Our bound is contrasted with the existing *upper bound* of $0.5n^{1.23}$, where $n = \sqrt{N}$, on the number of non-congruent solutions to the *Partial Digestion* (PD) problem. To our knowledge this is the first indicator that the PPD and the PD problems have different characteristics, despite their apparent similarities.

Supported by DOE grant DE-FG03-90ER60999, July 1990–January 1991.

P14
An Electronic Laboratory Notebook for Data Management in Physical Mapping

Debbie Nelson and J. W. Fickett
Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM 87545

We will demonstrate the current state of our computer-based information management system, the "Laboratory Notebook" database, which provides immediate and flexible access to experimental data. The database and associated software provide for data capture at the point of acquisition and allow uniform data management.

Compared to recording information in traditional paper notebooks, or in simple files, this system provides better data availability (immediate access to the most up-to-date version of all the data), better selectivity (easier to find a crucial fact in a large amount of data), and better flexibility (the data can be rearranged for different purposes).

Currently the mapping effort at Los Alamos uses the Laboratory Notebook to manage raw experimental data (e.g. clone fingerprints), processed data (e.g. deduced clone overlaps), data on samples and reagents (e.g. clones and restriction enzymes), and some project management data (e.g. noting which clones have been sent to collaborators). Recently we have extended the database to support STS data and the building of an integrated map.

The Laboratory Notebook is implemented in the Sybase relational database management system on Sun workstations. An intuitive forms-based interface has been developed which allows the average user to access the database without concern for details of data structures or command languages.

P15
The Reality of Probabilistic Fingerprint Analysis[1]

David O. Nelson, Elbert W. Branscomb, and Anthony V. Carrano
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94551

Many sources of error can arise in the automatic analysis of restriction fingerprints. First, algorithms must locate potential bands on a gel image or in a signal. Then other algorithms must decide which of the potential bands really are bands, and which are only random statistical noise, gel imperfections, biological debris, unusual concentrations of buffer, and the like. While some bands are obviously there, determining the reality of others can be extremely problematic. Thus bands can be missed, and other artifacts can be misidentified as bands. Finally, after deciding which are the "real" bands, other algorithms must assign each band a nominal fragment size, usually based on interpolating from a nearby size standard. Such interpolation is, by its very nature, inexact. Moreover, we must account for these error processes if we are to compute a realistic probability of observing the data we see under various overlap hypotheses.

During the past year, we have been improving and evaluating the effectiveness of our fingerprint analysis algorithms in the face of the realities mentioned above. We have focused on two issues: constructing a better decision algorithm for determining which putative peaks are real, and determining accurate error rates for the types of errors mentioned above.

To address the first issue, we have constructed an adaptive band selection algorithm which bases its decision about the reality of a band on estimates of the signal noise, plus the relative sizes and frequency of surrounding putative bands. This algorithm can detect smaller bands in areas of low signal, while concentrating on the larger bands in areas of strong signal.

To address the second issue, we studied fingerprint reproducibility by selecting a sample of inserts for multiple fingerprinting. We wanted to know how often detection and size errors occurred, so that we could include accurate error rates in our probability calculations. Of course, we also wanted some insight into what we could do to improve our fingerprinting and fingerprint processing algorithms.

We have drawn several conclusions from our analysis:

- Determining the posterior overlap probability between repeat fingerprints of the same cosmid provides a valuable measure of fingerprint reproducibility.
- In our method, non-reproducibility defined in this way is dominated by the random appearance and disappearance of bands, local signal nonlinearities, and to a much lesser extent, local differences in elution times.
- Fingerprints of DNA tagged with the yellow fluorescent label called "Tam" seem to be especially sensitive to artifactual peak generation. This may be due to the higher concentrations required to obtain adequate signal-to-noise ratios with this dye.
- We can effectively filter out many such problematic fingerprints by algorithms which assess the overall quality of the fingerprint.

[1]This work is performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. W-7405-ENG-48.

P16

Data Control and Security Issues for Distributed Genome Databases

Tom Slezak and Elbert Branscomb
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

Biological data has traditionally been stored in private databases, inaccessible to other researchers, or in various centralized repositories that provide dial-up access or update subscription services. It is becoming clear that such unrestricted private retention of large data sets generated with public funds should no longer be acceptable for genome grants. Centralized repository databases play many important roles for biological researchers, yet lack the immediacy of access to "fresh" data that is desired for collaborative research and community support. Recent advances in networking and databases permit a 3rd option: cooperating individual databases that function as a larger "distributed" database yet maintain complete local autonomy.

Some modern commercial relational databases provide true server/client access, which opens the possibility of allowing collaborators to have direct access to each other's data, under complete control of the owner. All details of access to remote database(s) can be hidden from end users with suitable front-ends. Issues of data security are in general handled trivially by the database package (by allowing read-only access to specified tables/views to external collaborators). Data control issues are more complex (e.g., if we collaborate with A and B by sending them clones/probes and put their results into our database, if both A and B are external users of our database they may not want the other to be able to view "their" data until they publish.) One suggested solution is to place a six-month hold on data before it could be seen by non-collaborators.

Without arguing the merits or ethics of this approach, we have decided to demonstrate the technical feasibility of multi-collaborator and distributed genome databases. Experiments were conducted with the help of Debra Nelson of LANL that proved the ease of use of allowing controlled access to external databases over the Internet. Other experiments were done at LLNL to verify that the 6-month "data hold" concept was feasible, at the cost of adding owner and timestamp fields to any tables that might contain "proprietary" data.

We view this means of data sharing as an important tool for communities of tightly-coupled collaborating researchers willing to work with external databases at a fairly low level. We recognize that objections may be raised by sites that have databases from different vendors, and note that Sybase front-end tools are relatively inexpensive for genome researchers. In addition, the LLNL C-language interface library was designed to be portable to other relational databases and would make it easy to "import" data from a collaborator's "foreign" database, or to do cross-database pseudo-join queries.

We conclude that multi-collaborator and distributed genome databases are technically feasible, necessary, and worth the overhead in data storage and database administration.

Given Internet access and Sybase front-end tools any collaborator can easily share our data in a tightly-controlled fashion that does not put an undue burden on the collaborator. Collaborators who wish to maintain their own Sybase database(s) can grant similar privileges to us, under their complete control. Work in progress with Johns Hopkins will allow GDB to access our physical mapping data using these methods.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract No. W-7405-ENG-48.

P17

Contig Assembly Program (CA)

C. A. Soderlund, P. Shanmugam, and C. A. Fields
Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003-0001

The Contig Assembly program (CA) solves the following problem: given multiple clones, each of which is made up of multiple fragments, determine if there exists one or more combinations of the fragments, such that the clones may be arranged in a contiguous line. CA works by determining overlaps, and is intended for use with total digest data. List all possible solutions.

Problems arise due to: 1) matching fragments do not always have exactly the same length, 2) unique fragments may have the same length, and 3) fragments may be missing. Due to these problems, CA often cannot find a solution even when one exists. To solve this problem, CA is used interactively, as follows: 1) CA is run with pairs of clones until a good overlap is found. 2) A new clone can be displayed to see the relation of its fragments to the existing contig. 3) If the new clone overlaps the existing clone, but has one of the problems stated above, the clone can be edited to correct the problem and then added to the contig.

Our current work concentrates on: 1) finding the "best" solution regardless of errors, 2) using STS data to resolve inconsistencies, and 3) allowing weights to influence what partial solutions are pursued.

P18

New Features in gm, Version 2.0

C. A. Soderlund, P. Shanmugam, and C. A. Fields
Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003-0001

Version 1.0 of the **gm** automated sequence analysis system was released in January, 1990, and has now been installed at several dozen laboratories worldwide. The **gm** system predicts the exon-intron organization of genes from genomic DNA sequence data, and displays the resulting exon maps and predicted amino-acid sequences via a graphic user interface (Fields and Soderlund, *CABIOS* 6 (1990) 263-270). It is designed to support incremental, exploratory analysis of new sequences as they are obtained, and is intended for use as a laboratory tool. We have spent the last year testing **gm** with sequences of known genes, using **gm** to analyze new sequences obtained by collaborators, and designing and implementing new functions. The new release, **gm** version 2.0, includes the following new features.

- Partial 3′-end cDNA data can be used to initiate the exon maps. This allows maps consistent with a partial cDNA to be examined either alone, or together with all other possible maps. By predicting only exons that consistently extend a known cDNA, **gm** can be used to efficiently design PCR primers for amplification of sequences from a total cDNA library.

- Predicted exon maps are now ranked in order of increasing protein-coding capacity. The user can choose to view only the highest-ranked nonoverlapping maps, each of which has the maximal coding capacity for the region that it spans. This procedure generates the maps most likely to produce hits in protein database searches, while greatly decreasing the total number of maps that the user must examine.

- Functions for displaying microrestriction maps, STS locations, cDNAs, and repetitive DNA elements at the same scale as the predicted exon maps have been added to the graphic interface. This allows predicted genes to be quickly aligned with physical maps, and facilitates the selection of restriction fragments to be used in probes of Northern blots.

- The graphic interface supports multiple system runs with different parameter settings. This facilitates use of **gm** as an exploratory analysis tool. It also allows **gm** to serve as an interface for displaying exon maps of known genes together with cDNA and physical mapping data. **menu** now includes a function that builds exon maps from known exon coordinates.

Our current effort is focused on developing functions to allow use of arbitrary cDNA data to initiate exon map construction, and on developing and testing improved compositional analysis tools. A version of **gm** incorporating these new tools is expected by early summer.

P19

Recent Enhancements to the LLNL Contig and Database Browser

Mark C. Wagner and Tom Slezak
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

We have developed a highly-automated system for doing DNA fingerprinting from cosmid clones. Output from the fully-automatic contig contruction and definition of minimal spanning paths is too complex and voluminous to be readily understood on paper. We have developed a graphical contig browser to assist in comprehending and manipulating this physical map data. Providing a visual representation of the data permits a better understanding of the overlap confidence relationships between the individual cosmid clones that comprise each contig.

The Contig and Database Browser has undergone extensive modification since the last Workshop in November 1989. Our current version is in daily use helping us to analyze the over 8,000 human chromosome cosmid clones fingerprinted and/or mapped to date. We have met our goals of typing the browser to a relational database system. We have written a database access library to minimize our dependency on any particular database, and to allow the ability to access both local and remote database simultaneously.

Other features added include the capability to view and manipulate two contigs simultaneously as well as the ability to analyze the underlying fingerprint waveform data from two or more cosmid clones. This makes it possible for the biologist to look at the entire history of the cosmid clone from beginning to end, without resorting to other tools. Attributes of cosmid clones are now also available to the biologist at the click of a button.

This tool is our model for a generalized graphical database browser that will allow ready access and visual display of all data generated on the Human Genome project. Work in progress will permit the browser to display all data objects stored in our database using a flexible object-based mechanism, incorporate externally-written display codes, support user query synthesis, and enable joins to be made between tables on local and remote database servers.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract No. W-7405-ENG-48.

P20

A Rapid General Method for Physically Mapping an Internal Sequence with Respect to the Ends of a DNA Molecule

J. C. Game[1], M. Bell[1], J. S. King,[2] and R. K. Mortimer[1, 2]
[1]Division of Cellular and Molecular Biology, Lawrence Berkeley Laboratory, Berkeley, CA 94720
[2]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720

We have used yeast chromosomes to test a method of mapping the physical position of cloned genes with respect to the ends of DNA molecules. Pulsed-field gels and Southern hybridization are employed, and information is obtained from the smear of molecular fragments that run below a band when the DNA comprising that band is randomly broken at low frequency. The distance between a cloned sequence and the nearest free end of a molecule can be determined by probing gels containing such randomly broken molecules. This is easily understood by considering molecules broken exactly once. Clearly, amongst the fragments of such molecules, none shorter than the distance between a gene and the nearest end of a molecule can contain that gene. When DNA break-frequency is low (<1 break/molecule), most fragments will arise from once-broken molecules. Hence, a given probe will mainly identify fragments extending downwards only to a size representing the distance of the probed gene from the nearest end of the molecule. Furthermore, a two-fold increase in the concentration of identified fragments is expected at and above a position corresponding to the distance from the probed gene to the more distant end of the molecule. This is because there are now two positions in the molecule where breakage will lead to fragments of a specific size, compared to one position for smaller fragments.

We have found that these two "threshold" changes in intensity can easily be observed in the smear patterns of DNA broken by X-rays. They correspond in position to the distances of the probed sequence from each end of the molecule, for several genes tested. We have also considered the probed fragment distribution expected from molecules broken more than once. We find that the expected distribution arising from molecules with two to several breaks shows peaks at size positions corresponding to the distance of the gene from the ends of the molecule. The peak representing the distance to the nearest end is strongest, and it diminishes more rapidly with increasing break frequency for genes towards the center of molecules than for loci near the ends. In practice, the peaks allow useful assessment of smears resulting from higher levels of breakage than those derived almost only from once broken molecules. We investigated the effect of increasing break frequency on the observed fragment distribution using X-rays, and calculated the expected distribution. We are developing protocols to optimize this physical mapping method both for intact chromosomes, and for restriction digested DNA as a means of locating genes by distance from restriction sites. We have mapped with this method *APN1*, a yeast gene of previously unknown position. Genetic crosses are currently being performed to confirm the physical mapping data. We are exploring the possibility of using this method to locate human genes that lie within 10 mb of a telomere. Also, we hope to use it to locate human genes on large restriction fragments and on YACs.

Supported by NIH Grant GM 30990 and DOE contract DE-ACO3-765F00098.

P21

Cloning and Characterization of Human Chromosome 21 YACs

Jeffrey C. Gingrich, Steven R. Lowry, Wen-Lin Kuo,* and Joe W. Gray*
Human Genome Center, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
*Biomedical Science Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

As part of the molecular biology efforts at the LBL Human Genome Center we have begun the cloning of human chromosome 21 DNA fragments into YACs. The restriction enzyme *Eag*I cuts human DNA into fragments from ~50 kb to ~500 kb and the DNA is compatible with the *Not*I cloning site of pYAC5. Our source of DNA is the mouse-human chromosome 21 hybrid cell line WAV-17 of which approximately 2% of the DNA is human.

Approximately 5,000 *Eag*I YAC transformants were arrayed into microtiter plates. Potential human YACs were first identified by colony hybridization using a cloned Alu repeat sequence, BLUR-8, as the probe. The YACs range in size from 40 kb to 500 kb, with an average size of ~100 kb. This size distribution reflects both the size range of the *Eag*I digested DNA and the inherent higher cloning efficiency of smaller DNA fragments in YACs. Further cloning experiments are in progress utilizing DNA size fractionated on pulse field gels to recover larger clones.

Approximately 70% of the human YACs show amplification using a series of PCR primers made against a consensus Alu sequence. We have developed new approaches to regional assignment of the YACs using the inter-Alu PCR products as probes both for Southern and for fluorescence in-situ hybridization experiments. Regional assignment of those YACs which do not amplify by inter-Alu PCR, and confirmation of the previously determined regional assignments using inter-Alu PCR products, is being done using biotinylated total yeast DNA as probes for fluorescence in-situ hybridization.

P22

Chromosome Region-Specific Libraries for Human Genome Analysis

Fa-Ten Kao and Jing-Wei Yu
Eleanor Roosevelt Institute for Cancer Research, and Department of Biochemistry, Biophysics, and Genetics, University of Colorado Health Sciences Center, Denver, CO 80206
(303) 333-4515, FTS (303) 333-8423

Molecular analysis and fine structure mapping of the human genome require isolation of large numbers of DNA probes from defined regions of the chromosomes. A direct approach is to use chromosome microdissection to remove physically the chromosomal region of interest and to clone the minute dissected material by microcloning. We propose to develop expertise in this micro-technology and to apply it to the construction of region-specific libraries for human genome analysis.

In our initial studies, 30 chromosomes of human chromosome 21 were microdissected using de Fonbrune micromanipulator. The dissected chromosomes were collected in a microdrop, treated with proteinase K, extracted with phenol, cleaved with MboI, and ligated to MboI linker-adaptor. All these steps were performed in nanoliter volumes under the microscope. The ligated material was amplified by PCR and cloned into pUC19. Highly stringent precautions were enforced to minimize extraneous DNA unintentionally introduced into these procedures. After cloning, 700,000 recombinant microclones were obtained. Colony hybridization showed that 42% of the clones contained repetitive sequences and 58% contained unique sequences. The insert sizes ranged from 50 to 1100 bp (mean 416 bp). Southern blot analysis confirmed that these microclones are of human origin and chromosome 21 specific. In addition, we used 4 microclones containing unique sequences to screen the human YAC library from Washington University and identified 5 positive YAC clones with inserts ranging between 50-360 kb. This indicates that it is feasible to use microclones with very small inserts to expand to a large genomic region for molecular genome analysis and for contig construction. Furthermore, the small insert size in the microclones is convenient for sequencing to identify STS as genomic landmarks. Finally, we used unique sequence microclones to screen a human liver cDNA library and isolated at least 4 expressed gene sequences among 100 microclones so far tested.

Physical mapping and sequencing of the human genome requires development of many new molecular approaches and methodologies. The chromosome microtechnology described here appears to offer an additional powerful methodology to achieve these goals. We will further develop expertise in this technology and apply it to human chromosomes 2 and 5 to construct various region-specific libraries for use in physical mapping by the DOE National Laboratories. If it proves promising, applications to other specific chromosomes will be warranted.

Abstract presented in the DOE Human Genome Workshop for Grantees and Contractors, Feb. 17-19, 1991, Santa Fe, NM.

P23

Linear Amplification DNA Sequencing Directly from Single Phage Plaques and Bacterial Colonies

B. Rajendra Krishnan, Robert W. Blakesley,* and Douglas E. Berg
Department of Molecular Microbiology, Washington University Medical School, St. Louis, MO 63110
*Life Technologies, Inc., Gaithersburg, MD 20898

Increasing the speed of sequencing short DNA segments is valuable for at least two types of oft-performed experiments: (i) analyses of gene structure and function entailing identification of base pair changes generated by localized mutagenesis; and (ii) specification of unique segments to serve as sequence tagged sites (STS) for use in genome mapping efforts. We report here a method of sequencing directly from phage plaques or bacterial colonies that is a labor-saving modification of the linear amplification method developed for partially purified DNAs by Murray (NAR 17:8889, 1989) and M. Craxton (submitted). Our protocol is amenable to automation and is more efficient than methods requiring prior PCR amplification when sequences of only a few hundred bp are needed.

In brief, λ or M13 plaques, or part of a bacterial colony containing a multicopy plasmid, such as pSPORT, are suspended in distilled water, and aliquots are then used directly for dideoxy chain termination sequencing with ^{32}P-end labelled primers, Taq polymerase, and 30 cycles of denaturation (95°C), annealing (temperature chosen based on length and GC content of primer), and extension (72°C). More than 200 bp of sequence were obtained from phage and plasmid clones with a variety of primers. The signals obtained with pSPORT or pBluescript-containing colonies or M13 plaques were more intense than those with λ plaques. The signals obtained with colonies stored for one or two weeks at 4°C were equivalent to those obtained with freshly grown colonies.

In conclusion, we anticipate that linear amplification sequencing directly from single phage plaques or bacterial colonies will prove useful in analyses of genome structure, sequence and function in diverse organisms.

P24

Assembly and Analysis of Contigs in the Carcinoembryonic Gene Family Region on Human Chromosome 19

Anne Olsen, Katherine Tynan, Harvey Mohrenweiser, Lori Johnson, Alex Copeland,
Brigitte Brandriff, Elbert Branscomb, Pieter de Jong, and Anthony Carrano
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory,
Livermore, CA 94550

The carcinoembryonic antigen (CEA)-like proteins are encoded by a large family of closely related genes that are members of the immunoglobulin superfamily. The CEA family can be divided into the CEA subgroup and the pregnancy specific glycoprotein (PSG) subgroup. Several different members of both subgroups have been mapped to chromosome 19q13.1–13.2. We have used a probe for the constant domain of the CEA coding sequence to identify CEA-positive cosmids in a human chromosome 19 library. Hybridization with this probe identified 238 positive clones in a 6–8X coverage of the chromosome, suggesting the existence of 30–40 CEA-related genes. As part of our effort to generate an overlapping cosmid map of chromosome 19, we have analyzed over 8000 chromosome 19 cosmids, including 228 of the CEA-positive cosmids, by a high resolution restriction fragment fingerprinting technique. This fingerprinting strategy has assembled 197 of the CEA-positive cosmids, together with 103 additional cosmids, into nine significant contigs consisting of 18 to 55 cosmids per contig. Hybridization with end-specific probes from cosmids at the ends of contigs has resulted in the joining of two different pairs of contigs, thus reducing the total number of contigs to seven. The minimum spanning paths of these contigs range from 85 to 220 kb and cover an estimated total distance of 1.2 Mb. Screening with probes for specific members of the CEA and PSG subgroups has assigned three of the contigs to the CEA subgroup and four contigs to the PSG subgroup. One of the CEA subgroup contigs with a 12-member spanning path has been analyzed by restriction digestion. The total length of this contig is 175 kb, with four distinct sites of hybridization to the CEA constant domain probe, each separated by about 20–50 kb. This indicates that several of these genes are tightly linked over relatively short stretches of DNA. Fluorescence in situ hybridization to decondensed sperm pronuclei (abstract by Brandriff et al.) is being used to estimate the total distance covered by the CEA family and to order and orient the seven contigs established by fingerprinting. Hybridization techniques employing YACs (abstract by Alegria et al.) and cosmid end probes are being used to extend and merge the established contigs to obtain closure in this region.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract No. W-7405-ENG-48.

P25

Monochromosomal Rodent/Human Hybrids Containing Dominantly Marked Chromosomes for Gene Mapping and Genome Fractionation

Mohinderjit S. Sidhu, Arbansjit K. Sandhu, Fan Chen, Beaula Helen, and Raghbir S. Athwal
Department of Microbiology and Molecular Genetics, New Jersey Medical School, Newark,
NJ 07103

We are developing rodent/human hybrid cell lines each containing a single different Ecogpt marked human chromosome (Monochromosomal hybrids). The presence of the Ecogpt gene in the human chromosome affords stable retention of the chromosomes by selection in a medium containing mycophenolic acid and xanthine (MX media). Experimental approach for producing these hybrids involves tagging of the chromosomes in normal human fibroblasts or lymphoblastoid cells by infection with a retroviral vector carrying the Ecogpt gene. Following infection, clones of human cells isolated by selection in MX media are analysed for the identity of Ecogpt marked chromosome. This chromosome is identified by PCR based cloning of DNA flanking the site of vector integration and Southern hybridization to DNA's from a panel of hybrid cell lines. Using this approach, we anticipate to generate 24 different cell lines each containing a different marked chromosome.

In view of limited life span of normal human cells, marked human chromosome is perpetuated by fusing with mouse A9 cells. The human chromosome is then further transferred to mouse or Chinese hamster cells by the procedure of microcell fusion. Microcell hybrids isolated by selection in MX medium are analysed to confirm the identity and integrity of the transferred chromosome. The ultimate objective of this project is to generate two sets of monochromosomal hybrids, one in mouse cells and the other in Chinese hamster cells. We have already produced monochromosomal hybrids for chromosomes 2, 5, 6, 7, 9, 13, 15 and 17.

P26

A Two-Dimensional YAC Pooling Strategy for Rapid Screening Via STS and Alu-PCR Methods

Michelle Alegria, Chris Amemiya, Charalampos Aslanidis, Chira Chen, Jeff Gingrich,* Julia Nikolic,* and Pieter de Jong
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550
*Human Genome Center, Lawrence Berkeley Laboratory, Berkeley, CA 94270

The STS-based PCR screening method of Green and Olson (*PNAS* 87:1213-1217) has become the routine method for isolation of YAC clones. The method is rather tedious and time-consuming, particularly since the final step relies on colony hybridization to YAC arrays. While the colony hybridization step may be supplanted via a simple plate-PCR method (Kwiatkowski *et al.*, *Nucl Acids Res.* 18:7191-7192), the overall technique is still comparatively labor-intensive since it requires three successive rounds of screening and time consuming since it includes growing up all clones for the YAC microtiter dishes identified in the 2nd screening step. We have developed a modification of this method that relies on a two-dimensional YAC pooling scheme in which candidate clones can generally be identified after two successive rounds of PCR without the need to access the microtiter dishes until the candidate clone is identified. The 55,000-clone genomic YAC library of Olson *et al.* (Washington University) was divided into six 96-plate sets, each set representing *ca.* 0.8× coverage of the human genome. Two different pooling schemes ("dimensions") were completed for replicates of the respective sets. The pooling schemes (see Amemiya *et al.*) were: 1st dimension -- each pool representing all clones from a single 96-well dish ; and 2nd dimension -- each pool representing a constant well position for 96 plates. A total of 1152 pools was thus required to represent each dimension: 6 sets × 2 dimensions × 96. To facilitate PCR screening of the library, 48 superpools were prepared for each dimension, with each pool representing 12 regular pools. The first round of STS-PCR screening is done on both the 1st and 2nd dimension superpools. Positive superpools are noted, and STS-PCR subsequently performed on the 12 pools corresponding to each positive superpool. Each clone in the library is uniquely defined by a combination of a 1st and 2nd dimension pool. We have used this screening method for isolating various chromosome-19-specific YACs, including about 12 YACs containing gene sequences for CEA (carcinoembryonic antigen), and several YACs mapping distal and proximal to the presumptive locus for DM (myotonic dystrophy; see Aslanidis *et al.*). In addition, we have successfully screened the superpools and pools by Southern hybridization using *Alu*-PCR probes derived from other YACs, from cosmids or by subcloning from hybrid cell lines.

Work performed by the Lawrence Livermore National Laboratory under the auspices of the U.S. Department of Energy under contract No. W-7405-ENG-48.

P27

A Novel Contig Mapping Strategy (MPAP) Based on Multi-Dimensional Pooling and Alu-PCR

Chris Amemiya, Michelle Alegria, Jennifer Alleman, Charalampos Aslanidis, Chira Chen, Alex Copeland, and Pieter de Jong
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

Our previous mapping efforts have largely focused on the construction of a chromosome-19 contig map using a fluorescence-based fingerprinting method (see Carrano *et al.*). A new mapping strategy, heretofore referred to as MPAP (Multi-dimensional Pooling-Alu-PCR) is currently being implemented in order to confirm, expand and link cosmid contigs. This approach relies on Alu-PCR and a five-"dimensional" pooling scheme of our 10,000-clone arrayed chromosome-19-specific cosmid library. Five replicates of the cosmid library were made and individual clones were combined in different pooling schemes ("dimensions"), such that each clone in the library is uniquely identified by a combination of pools from respective dimensions. The pooling schemes were: 1st dimension -- per microtiter dish (108 total pools, 96 clones/pool); 2nd dimension -- per well position; 3rd dimension -- per well position with positive offset; 4th dimension -- per well position with negative offset; 5th dimension -- similar to 4th dimension but with a slightly different offsetting scheme. Dimensions 2 through 5 are each represented by 96 total pools with 108 clones/pool. Alu-PCR was performed on all cosmid pools and the resulting products were arrayed in 96-well microtiter dishes according to respective pool number. The Alu-PCR products were then spotted on a membrane for hybridization detection of positive pools using chromosome-19-specific Alu-PCR probes, e.g., from cosmids, YACs or region-specific clones. We have demonstrated that we can readily detect positive pools well above background levels. By scoring the hybridization patterns for the different dimensions, we can deduce which clones are putatively responsible for the observed hybridization signals. Several cosmids and YACs have been successfully used to generate Alu-PCR probes for detection of overlapping cosmid clones. In addition,12 region-specific (19q13.2-13.3) Alu-PCR "coincidence clones" have been assigned to corresponding contigs. The results show that MPAP is a viable and rapid cosmid contig mapping method. To facilitate "closure" of the chromosome-19 contig map, we are enlisting YACs as "linking" tools (see De Jong *et al.*, Aslanidis *et al.*, Alegria *et al.*) and have initiated pilot experiments to test the feasibility of using MPAP on YACs. A saturation mapping approach using cloned chromosome-specific Alu-PCR products as hybridization probes is currently being implemented.

Work performed by the Lawrence Livermore National Laboratory under the auspices of the U.S. Department of Energy under contract No. W-7405-ENG-48.

P28

Progress Report on the Physical and Linkage Map of Human Chromosome 21

S. E. Antonarakis, P. A. Hieter, M. K. McCormick, A. C. Warren,* M. Kalaitsidaki,*
A. Chakravarti,** M. Petersen,*** and S. Slaugenhaupt**
School of Medicine, *Department of Psychiatry and Center for Medical Genetics, Johns Hopkins University, Baltimore, MD 21205
**University of Pittsburgh, Pittsburgh, PA 15261
***John F. Kennedy Institute, Glostrup, Denmark

The major goal of this laboratory is to contribute to the mapping of human chromosome 21. The progress to date is as follows:

1. Genetic map. A total of 31 DNA markers have been fully genotyped in CEPH pedigrees and a multipoint map has been constructed that spans approximately 130 cM. About 60% of recombination occurs in the distal 1/5 of the long arm and the female map is longer than the male map. The average distance between markers is 6 cM. The map contains 12 genes and 7 loci markers with PCR-based polymorphisms. A large number of markers based on short repeat polymorphisms are under development and will be added to this map.

2. Physical map. A contig map using overlapping yeast artificial (YACs) and cosmids has been initiated. The project involves:

   2.1 Development of STS for chromosome 21. These STS are from (i) known and well mapped probes, (ii) plasmid inserts from the J. Gray chromosome 21 library that have been screened with short sequence repeats, (iii) ends of YACs from the Johns Hopkins collection, (iv) ends of cosmids from the H. Lerach chromosome 21 library.

   2.2 Screening of YACs with STS. The YACs available for screening are (i) The Johns Hopkins collection of chromosome 21 YACs made in vectors pJS97 and pJS98. The average size insert is about 300 kb and 74 YACs have been characterized. The cloning was from cell line WAV-17 that contains only human chromosome 21 and from flow-sorted chromosome 21, (ii) the Los Alamos chromosome 21 YAC collection, cloned in pJS97 and pJS98 after flow sorting chromosome 21, (iii) the CEPH YAC library in pYAC4 that covers 7X the human genome, (iv) the ICI YAC library in pYAC4 that covers 3X the human genome.

   2.3 Screening cDNA libraries for expressed sequences on chromosome 21. The screening is performed using (i) Alu-PCR amplification products from cell line WAV-17 and YACs or (ii) whole YACs and cosmids after suppression of repeated sequences by CotI DNA. A small number of cDNAs have already been cloned and partially characterized.

P29

Isolation of YACs and Cosmids from 19q13.3

Charalampos Aslanidis, Chris Amemiya, Michelle Alegria, Jennifer Alleman, Chira Chen, Gert Jansen,* Gary Shutler,** and Pieter J. de Jong
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550
*Department of Cell Biology, University of Nijmegen, Nijmegen, Netherlands
**Department of Microbiology and Immunology, University of Ottawa, Ottawa, Canada

The q13.3 region of human chromosome 19 contains the elusive myotonic dystrophy locus (DM). As a result of the high medical interest, many physical and genetic markers have been obtained from the DM region. Distal and proximal flanking markers for DM span an area of approximately 500–1000 kbp. On the proximal side of DM, we have walked in the chromosome 19-specific cosmid library. Eleven walk steps have led to a 350 kbp cosmid contig. A sizeable portion of this contig has been confirmed independently by fluorescent fingerprinting of random chromosome 19 cosmids. The fingerprint analysis has also extended the contig in the centromeric direction (from ERCC1). The estimated size of the entire proximal contig is 500 kbp. For the flanking marker distal to DM ("X75"), a small (70 kbp) cosmid contig consisting of 4 cosmids has been established by hybridization with the corresponding probe. To link the distal and proximal cosmid contigs, we have isolated YACs corresponding to both contigs. Screening of the YACs is being performed by two approaches: either by applying STSs or by hybridization using *Alu*-PCR probes isolated from cosmids. These approaches are outlined on another poster (Alegria *et al.*). For the **proximal** contig, a 360 kbp YAC was isolated by screening with STS primers. The 360 kbp YAC contains nearly the complete cosmid walk but does not appear to expand the contig. Further YACs are presently being isolated by hybridizing with a 470 bp *Alu*-PCR probe isolated from a cosmid from the end of the cosmid walk. To expand the **distal** contig, a PCR primer pair from the vicinity of the X75 marker was used to detect two YACs (240 and 220 kbp). Probes from all YAC insert ends were generated by vector/*Alu*-PCR. The products were then cloned and sequenced. The orientation of the two distal YACs was determined using the end probes on hybrid cell panels containing hybrids with breakpoints in this region. To identify cosmids corresponding to the YACs, we applied the Multidimensional Pooling/*Alu*-PCR procedure (see poster by Amemiya *et al.*). Further distal YACs have been isolated using STSs corresponding to the subcloned YAC insert termini.

Work performed by the Lawrence Livermore National Laboratory under the auspices of the U.S. Department of Energy under contract No.W-7405-ENG-48.

P30

Refined Linkage and Physical Mapping of Chromosome X Genetic Markers and Pilot Studies for the Isolation of X-Specific cDNAs from Organ Specific Libraries

David F. Barker, Pamela R. Fain, and Arnold R. Oliphant
Division of Genetic Epidemiology, University of Utah, Salt Lake City, UT 84112

We have previously reported the isolation of 80 RFLP markers for the X chromosome and the mapping of these markers to 25 distinct intervals with an X somatic breakpoint panel. Genetic mapping of the same markers has also been pursued in the CEPH linkage reference family panel. With the addition of new breakpoints, the interval map now defines

35 distinct locations. A genetic map of 95 X-specific markers has been constructed with 50 markers from our set and data from 45 markers in the CEPH database.

Correlation of the genetic and breakpoint interval maps with the chromosomal and the overlapping clone maps is in progress. Dr. David Ward at Yale University is using probes from our collection for fluorescent in situ hybridization to provide a precise cytogenetic localization. Dr. David Schlessinger has assigned our probes from the xq27-q28 region to YAC contigs which he has developed in that region. Several laboratories, including our own, are developing YAC contigs in the vicinity of specific disease genes of interest, using probes from this collection.

Pilot experiments are in progress to define conditions for utilizing DNA prepared from the total LAOXNL01 library as a hybrid selective reagent for the enrichment of X-specific clones from organ-specific libraries, including a fetal kidney library and a fetal retinal library. The status of these will be presented.

Supported by DE-FG02-88ER60689.

P31

Current State of the Physical Map of Human Chromosome 16

D. F. Callen, L. Z. Chen, J. Nancarrow, S. A. Whitmore, S. Apostolou, A. D. Thompson,
S. A. Lane, R. L. Stallings,* C. E. Hildebrand,* P. G. Harris,** and G. R. Sutherland
Department of Cytogenetics and Molecular Genetics, Adelaide Children's Hospital, North Adelaide, SA, Australia
*Los Alamos National Laboratory, Los Alamos, NM 87545
**Institute of Molecular Medicine, John Radcliffe Hospital, Oxford, United Kingdom

An extensive mouse/human hybrid cell panel of human chromosome 16 has been constructed (see poster Callen, Lane *et al.*). The breakpoints in these hybrids, together with the four major fragile sites on this chromosome, have been ordered by mapping gene and anonymous DNA probes. A PCR approach to physical mapping has now been developed by multiplexing a series of STS markers (see poster Richards *et al.*) and by mapping of AC-repeat microsatellite sequences (see poster Mulley *et al.*). At this stage chromosome 16 has been divided into thirty-nine regions by mapping 184 markers. The short arm of chromosome 16 has been divided into twenty well distributed regions. The nineteen regions on the long arm are clustered with ten hybrid breakpoints being located in q22.1. However, there are several breakpoints in q12 to q13 and q23 to q24 which have yet to be mapped in detail. In some cases breakpoints are apparently coincident since probes have not yet been mapped between them.

Either within our laboratory, or in collaboration with other laboratories, approaches are in progress towards cloning particular regions. These include fragile sites, Batten disease and various areas involved in malignancy.

P32

A Somatic Cell Hybrid Panel for Mapping Human Chromosome 16

D. F. Callen, S. A. Lane, H. Eyre, E. Baker, and G. R. Sutherland
Department of Cytogenetics and Molecular Genetics, Adelaide Children's Hospital, North Adelaide, SA, Australia

Forty-seven mouse/human hybrids have now been constructed containing various portions of chromosome 16 by utilising selection of the gene APRT located at the distal tip of the long arm. The majority of the human cell lines used as parents in these fusions contained balanced reciprocal translocations identified in cytogenetics laboratories. In addition there are three interstitial deletions (CY160, CY130 and CY125), an interchromosomal insertion (CY180) and a complex translocation associated with leukemic cells (CY105). In general, these hybrids are relatively stable with the retained portion of 16 being intact. However, three derivative hybrid lines were isolated containing portions of chromosome 16 resulting from mouse/human rearrangements or chromosome 16 rearrangements (CY180A, CY13A and CY18A).

The breakpoints contained in these hybrids have been ordered by mapping probes (see poster Callen, Chen *et al.*). Since the majority of these hybrids are derived from reciprocal translocations they contain portions of chromosomes other than 16. This provides a useful resource for physically mapping other chromosomes. For example, CY5 and CY6, which contain breakpoints on chromosome 10, have been used for physical mapping on this chromosome by P. J. Goodfellow (personal communication).

P33

The Status of the Chromosome 19 Physical Map

A. V. Carrano, M. Alegria, J. Alleman, C. Amemiya, L. K. Ashworth, C. Aslanidis, B. Brandriff, E. W. Branscomb, L. Brown, C. Chen, M. Christensen, J. Combs, A. Copeland, P. J. de Jong, A. Fertitta, E. Garcia, L. A. Gordon, L. Johnson, C. Kwan, J. Lamerdin, H. Mohrenweiser, D. Nelson, A. Olsen, T. Slezak, B. Trask, K. Tynan, and M. Wagner
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

We have constructed a cosmid contig map of human chromosome 19 which we estimate spans about 70% of the chromosome. This foundation map currently consists of 720 contigs and was assembled by automated fluorescence-based fingerprinting of 8813 chromosome 19-specific cosmids. The validity of the map is being assessed by several methods including integration with the genetic map, fluorescence *in situ* hybridization (FISH), and restriction enzyme site mapping. Closure of the cosmid contig map is being accomplished by hybridization to YACs and other cosmids. Alu-PCR probes generated from reduced hybrids of chromosome 19 are being used to screen a YAC library for YACs associated with chromosomal regions. Alu-PCR products from these YACs are then hybridized to Alu-PCR products from our cosmids to identify cosmids corresponding to the YACS. In addition, unique sequence probes from YACs or cosmids are used to screen both a YAC and cosmid library for clones bridging the gaps in our contig map. Orientation of the contigs is provided by linkage of the contigs to genetic markers and by FISH to metaphase chromosomes, to somatic interphase nuclei, and to sperm pronuclei. A total of 47 probes (41 unique), representing genetic markers on chromosome 19, have been hybridized to our cosmid library. Thirty-three of the probes (70%) identified previously assembled contigs.

Of these 33, 24 probes each identified a single contig, 6 probes each identified two contigs, 1 probe identified three contigs, 1 probe identified 15 contigs, and 1 probe identified 74 contigs. The latter two probes were associated with a multigene family (CEA) and a minisatellite repeat element (pE670), respectively.

We have mapped 302 cosmids to metaphase chromosomes by FISH. Of these, 242 fall into a total of 105 contigs. Eighty-two of the 302 mapped cosmids are associated with one or more genetic markers. Multicolor hybridization of multiple cosmids from the same contig or chromosome region to metaphase chromosomes and interphase cells provides coarse map location and contig validation.

This approach has given us an initial order of contigs in the q13.2 region. Hybridization to pronuclei provides high resolution information on order, gene size (e.g. the CEA family), and gap size. Our physical map information is stored in a relational database and is accessible to end-users via SQL or graphical query. We are currently sequencing a region of chromosome 19 surrounding a DNA repair gene (ERCC1) and are beginning the selection, sequencing, and mapping of cDNAs for use as STSs.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract No. W-7405-ENG-48.

P34

Thermal Stability Mapping

Nashwa W. Gabra, Eric S. Schmitt, William J. Fripp, and Leonard S. Lerman
Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139

This mapping system determines the order and position of two types of markers along genomic DNA. One class of markers, low melting domains, are defined by regions of lower thermal stability than neighboring regions. The second class consists of all sites for which specifically hybridizing probes can be constructed. Data is obtained directly from operation with genomic DNA, without cloning or amplification. While restriction sites have no role in defining the map, there is full compatibility with restriction maps. The data are derived from the patterns of probe hybridization with two-dimensional denaturing gradient separation of random fragments of genomic DNA. Random fragmentation provides a DNA sample in which every short sequence (roughly 1/2 Kb) is represented in a broad distribution of fragment lengths with randomly distributed 3′ and 5′ ends. Fragments are separated in the first dimension by length and in the second by retardation according to the least stable domain in the denaturing gradient. The gradient separates molecules according to a hierarchy of the retarding domains into distinct families. These generate well-defined, characteristic two-dimensional patterns when subjected to probe hybridization.

We show the results of these procedures as applied to bacteriophage lambda DNA, the principles by which map distances are inferred from the patterns, and a preliminary lambda map. The data are also compared with the patterns expected on the basis of melting and electrophoretic theory.

P35

Automated Fluorescence-Based Restriction Fragment Analysis

J. Lamerdin, K. Corcoran,* P. E. Mayrand,* A. Olsen, K. Tynan, M. Kronick,* and A. V. Carrano
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550
*Applied Biosystems, Inc., Foster City, CA 94404

The accurate measurement of restriction fragments is an important component of most molecular biological applications. In particular, for the human genome initiative, one would ideally like to have a restriction map of the contigs that are generated for a chromosome or region. Restriction fragment analysis also allows for quality control; it provides clone insert length, confirmation of overlap between clones, and length of genomic region covered in contigs. We have adapted the chemistry procedures currently used for our cosmid fingerprinting to the development of six-cutter restriction maps for cosmids from our chromosome 19 contigs. The procedure was facilitated by the utilization of a prototype instrument called the "GENE SCANNER" which is now marketed by Applied Biosystems Inc. The system uses a horizontal agarose gel with laser excitation of fluorochrome-labeled fragments similar to the fluorescence-based DNA sequencer. Cosmids are digested with restriction enzymes and a fluorescent linker is simultaneously ligated to the 5′-overhang. Four fluorescent dyes can be utilized for the fluorochrome-linker. One of the dyes (ROX) is used to label the internal size standard in every lane, (λ+pSP64/BamHI) + (λ+pBR322/HindIII). The size standards used for cosmid analysis range from 600 to 27,527 bp and electrophoresis is performed in 0.8% agarose (FMC, SeaKem GTG). Other size standards can be used to measure fragment sizes down to =100 bp using 2% agarose (FMC, SeaPlaque). One or more cosmids, digested with a six-cutter and labeled with a fluorochrome, are loaded with the size standard in a single lane of the gel. Twenty-four lanes can be loaded simultaneously. The distance from the loading well to the laser excitation and read region is variable but we found that about 4 cm is optimal. Electrophoresis is at 130V (4.8V/cm) and is completed within 5 hrs. Data collection and analysis of fragment sizes are performed with Macintosh-based software. We have analyzed over 200 cosmids from our chromosome 19 contig map. EcoRI digests of our cosmids produce a 1.24 and 3.75 kbp vector fragment. The coefficient of variation (CV) on repeat measurements of these fragments is 1.5% and 0.5%, respectively. The CV on measurements of the same fragment sizes in overlapping cosmids averages about 1% over the range from 1–18 kbp. Because the dyes and fragments are measured at a fixed point in migration, the same gel can be used several times. In addition, the gel can be removed from the system, stained with ethidium bromide for confirmation of fragments, and the DNA transferred to membranes for Southern analysis. In addition to providing quality control information for our cosmid contigs, we are adapting these methods for the automated restriction fragment fingerprinting of PCR fragments and for the partial digest analysis of cosmids.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract No. W-7405-ENG-48.

P36

Identification of Reference (Framework Loci) Cosmids on the Contig Map of Human Chromosome 19

H. W. Mohrenweiser, K. M. Tynan, B. Brandriff, E. Branscomb, P. de Jong, B. Trask, and A. V. Carrano
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

The integration of the genetic/linkage and physical maps of the genome is a requirement for maximizing the utility of physical maps for future applications, including the isolation of genes associated with genetic diseases. This integration is best accomplished by identifying the cosmids associated with genetically mapped markers. Twelve highly informative, genetically well mapped markers have been defined as constituting an initial reference or framework map for human chromosome 19. Cosmids associated with 8 of the framework loci, as well as cosmids associated with 21 other genetically mapped (polymorphic) markers have been isolated from a chromosome 19 specific library. These cosmids, an average of ~6 per marker, along with >8000 additional cosmids from this library, have been fingerprinted (Carrano et al., Genomics 4:129,1989) for assembly into contigs. Further analysis of the cosmids associated with the framework locus D19S11 indicate that this compound polymorphic locus is actually 4 loci contained in 3 contigs spanning ~325 kb. Fluorescence in situ hybridization of these cosmids to interphase sperm pronuclei suggests that these markers span ~500 kb, consistent with the cosmid map size.

Each of the framework-associated contigs analyzed thus far maps to the appropriate chromosome region by *in situ* hybridization. PCR primers that yield a single, appropriately sized amplification product from genomic DNA have been synthesized for 4 of the markers. The PCR products are currently being used as probes for hybridization against genomic DNA blots as necessary to confirm their utility as STSs.

Sixty markers are necessary for developing an integrated genetic-physical map of chromosome 19 with markers spaced at 1.0±0.5 cM. At this point, cosmids associated with 18 appropriately spaced, polymorphic markers have been identified.

Work performed under auspices of the U.S. DOE by the Lawrence Livermore National Laboratory; contract No.W-7405-ENG-48.

P37

Progress on Making a Complete Restriction Map of Human Chromosome 21

Denan Wang, Jesus Sainz, Rafael Oliva,* Cassandra L. Smith, and Charles R. Cantor
Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, and Human Genome Center, Lawrence Berkeley Laboratory, Berkeley, CA 94720
*Grupo de Genetica Molecular, Facultad de Medicina, Universidad de Barcelona, Spain

We are constructing maps of human chromosome 21, the smallest human chromosome by several strategies. The initial approach involves generating large DNA fragments with infrequently cutting restriction enzymes like *Not* I, separating these fragments by pulsed-field gel electrophoresis (PFG) and locating specific fragments on the chromosome by hybridization with genetically mapped probes. In addition, partial digestion strategies are used to see neighboring fragments; linking clones are used to identify adjacent pieces of DNA, and cell line polymorphism link up is used to identify probes on the same or adjacent fragments. We have focused on first finishing the map in the q telomere region. About 21 single copy gene probes and 8 linking clone probes have been used to detect *Not* I fragments in band q22.3, the telomeric long arm Giemsa light band. We find that smaller *Not* I fragments occur preferentially in this region. Since *Not* I sites occur almost exclusively in G-C rich islands, this result suggests that band q22.3 is unusually rich in genes. Our results have produced extensive maps covering much of chromosome 21, but there are still gaps. We have developed a new strategy to fill these gaps. We are using inter Alu PCR to generate human-specific DNA probes from slices of PFG-fractionated *Not* I-digested hybrid cell DNA. Such probes can be hybridized to partially digested genomic DNA to identify the neighboring fragments.

DNA Sequencing/Instrumentation Database Needs

P38

Film-Reading Aids for Molecular Biology

J. B. Davidson
Instrumentation and Controls Division, Oak Ridge National Laboratory,* Oak Ridge, TN 37831

Two viewing aids which have been developed for reading sequencing, blotting and other films will be described and demonstrated:

The Un-Dimmer is a novel, hand-held viewer which improves the contrast of bands or spots that are invisible or nearly so above background when the conventional transmission light box is used.

The Un-Smiler is a specialized passive magnifier which can be used to rectify or "unsmile" inclined bands on DNA sequencing films. The bands can be made to appear essentially parallel to a fiducial line in the viewer as the viewer is moved over the film. An improved version will be shown.

Principles employed in both viewers can be used in scanning densitometers and image digitizers.

*Managed by Martin Marietta Energy Systems, Inc., for the U.S. Department of Energy under Contract DE-AC05-84OR21400.

P39

Development of a Laboratory Database for Physical Chromosome Mapping

Reece K. Hart and Glen A. Evans
Molecular Genetics Laboratory and Center for Human Genome Research, The Salk Institute for Biological Studies, La Jolla, CA 92138

The construction of physical maps of human chromosomes necessitates the development of new computer databases for the storage, retrieval and analysis of diverse types of mapping information. These databases will be most heavily used by molecular biologists with little interest or patience in computer systems having a demanding user interface. Our goal has been to develop a simple yet powerful database system for use in our chromosome 11 mapping project which satisfies four criteria: 1) ease of use, 2) powerful, 3) inexpensive and 4) entertaining. We recently completed the first phase in the development of an integrated database system for storage of diverse types of chromosome mapping information and utilized this for the analysis of information derived from the human chromosome 11 mapping project at the Salk Institute. This system runs on Macintosh computers using the Hypercard user interface as a shell for integrating several custom extensions written in the C programming language. This database can utilize and store information in the form of text, bit-mapped and object-oriented graphics, digitized sound and speech, and images obtained from a digitizing scanner or video camera. Currently, data on cosmids, yeast artificial chromosomes, Cosmid/YAC contigs, pulsed field gels, DNA sequences, PCR primer sets and sequence-tagged sites may be stored and analyzed using a graphics-oriented display. Key features of this database include 1) the ability to integrate data from a number of mapping strategies (cosmids, YACs, contigs, PFG and DNA sequence), 2) the ability to continually modify and evolve software while it functions as an operational database, 3) the presentation of a facile user interface requiring little or no training for operation, 4) the extensive use of digitized sound and speech in the user interface, and 5) the extremely short development, implementation and modification time due to the rapidity of programming in the Hypercard environment. Future developments will include implementation of the database on a SUN/4 multiuser network server using a Macintosh/Hypercard user interface identical to the one employed in the current operational database system.

P40

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

John R. Hartman
Computational Biosciences, Inc., Ann Arbor, MI 48106

As the growth of DNA and protein sequence databases continues to accelerate, exhaustive methods for searching them for homologies and biologically meaningful features have become increasingly impractical on traditional serial-architecture computers. Recently, considerable work has been done to implement these methods on supercomputers and high-end massively parallel processors, but the expense and poor accessibility of these machines place them out of reach of most molecular biologists. In Phase I of this project, several efficient algorithms for the exhaustive comparison and homology searching of macromolecular sequence data were implemented on a relatively inexpensive SIMD (Single Instruction/Multiple Data stream) parallel array computer. These were carefully and rigorously evaluated with respect to their correctness and performance behavior, and one in particular was found to offer impressive cost/performance advantages over functionally equivalent serial software. The system as presently configured consists of a Sun SPARC station 1+ host in addition to the parallel computer, which also includes an MC68020 processor that can be utilized concurrently. During Phase II, overall system throughput will be maximized via a detailed multiplex optimization strategy designed during Phase I. Also, an X-Window-based graphical user interface following the OpenLook specification will be designed and implemented, and significant new sequence analysis and data management functionality will be developed.

The success of this project will result in the commercial introduction of a parallel processing sequence analysis workstation with robust capabilities and unprecedented cost/performance characteristics. The availability of very high-performance sequence analysis capabilities at reasonable cost will facilitate a decentralized approach to the Human Genome Project and substantially improve the return realized on research dollars invested.

P41

Application of Storage Phosphor Image Plates to Autoradiography

W. F. Kolbe, W. H. Benner, J. M. Jaklevic, L. E. Sindelar, and J. Gingrich
Human Genome Center, Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720

We describe the use of photostimulable storage phosphor imaging plates as a substitute for X-ray film in the visualization of radioactive patterns. Previously published results have shown the technique to have superior sensitivity, linearity and dynamic range when compared to film. Using a commercial system (Molecular Dynamics, Sunnyvale, CA, model 300), we have employed artificially generated patterns labeled radioactively with ^{32}P ink to quantitatively evaluate the limitations of the technique for mapping and sequencing applications. Comparisons of image plate results with X-ray film images have been obtained for Southern blots and high-density dot-blot hybridization patterns. Advantages and disadvantages of the phosphor image plate approach will be discussed. In addition, the application of the method to imaging of extremely low activity level samples will be addressed.

P42

Human Genome Management Information System

Betty K. Mansfield, Anne E. Adamson, Denise K. Casey, K. Alicia Davidson, Roswitha T. Haas, Kathleen H. Mavournin, Elizabeth T. Owens, Judy M. Wyrick, Laura N. Yust, and John S. Wassom
Human Genome and Toxicology Group, Health and Safety Research Division, Oak Ridge National Laboratory,* Oak Ridge, TN 37831-6050
615/576-6669; Fax 615/574-9888; E-mail: "BKQ@ORNL.GOV"

The Human Genome Management Information System (HGMIS), cosponsored by the Department of Energy (DOE) and the National Institutes of Health (NIH), has roles in the international Human Genome Project to:

1. assist agencies that administer genome research in communicating issues relevant to the Human Genome Project to contractors and grantees and to the public and

2. provide a forum for exchange of information among individuals involved in genome research.

To fulfill these communications goals, HGMIS is producing a bimonthly newsletter, DOE Human Genome Program reports, an information database, and technical reports. HGMIS updates and maintains the mailing list database compiled for the human genome programs of both DOE and NIH. Additionally, HGMIS acts to orient and refer those persons seeking assistance to sources that can provide appropriate information. These documents/services are available to all persons upon request and provide both the interested scientist and lay person with information in this rapidly moving, multidisciplinary project.

- The newsletter, *Human Genome News* (ISSN #1050-6101), provides readers with technical and general interest articles, meeting reports, news items, funding announcements, and meeting and training calendars. Working in collaboration with the international Human Genome Organisation, HGMIS also reports international genome project news.

- The status of the DOE Human Genome Program is described in the *Human Genome 1989-90 Program Report,* which includes research highlights, narratives on major DOE research efforts, abstracts of research in progress, and figures and captions provided by investigators.

- The information database is being developed as a text management and userconferencing/communications mechanism. It contains text from program reports and newsletters, bibliographic data from both scientific and popular literature, and current awareness items.

- Technical reports will be commissioned and produced by HGMIS as recommended by the DOE Human Genome Program.

*Managed by Martin Marietta Energy Systems, Inc., under Contract No. DE-AC05-84OR21400, for the Department of Energy.

P43

GENCORE: An Automatic Genetic Database Cross-Correlator

Stanley M. Schwartz
CHI Systems Incorporated, Spring House, PA 19477

CHI Systems is developing a cross-referencing tool and intelligent interface intended to address the unique needs of users and developers of molecular biology databases. GENCORE (An Automatic Genetic Database Cross-Correlator) will automate the process of generating, implementing, and storing the results of cross-reference searches across sequence, mapping, crystallographic, and bibliographic databases (with incompatible formats). Based on Phase I SBIR research, CHI Systems has designed a generalized modular architecture for GENCORE that provides for independent and extensible access to multiple databases, whose DBMS's will be interconnected via a common translation, storage, analysis, and data-entry/editing tool. The design for GENCORE also incorporates a cognitively engineered, platform-independent user interface that will allow the specification of sequence and annotation search criteria to be simultaneously applied to single or multiple databases.

P44

The Computational Linguistics of the Genome

David B. Searls
Unisys Center for Advanced Information Technology, Paoli, PA 19301

Most current computational approaches to the analysis of biological sequences pay only lip service to the fact that DNA is a *language*, and thus amenable to methods of linguistic analysis that have been extensively studied in other contexts. While abstracted, hierarchical views of sequence information are emerging, in some cases from the AI field, there appears to be a need for a uniform, formally grounded system to better encompass the diverse functions currently performed by a wide variety of software. For this purpose, computational linguistics offers both an intensively studied declarative representation (formal grammars) and an equally well-developed procedural interpretation (parsing). Our results suggest that not only DNA is linguistically "interesting," but pattern-matching search and other forms of analysis may profit from such a linguistic approach.

The question then naturally arises as to where the language of DNA is situated relative to known language classes. Pattern-matching algorithms now used for DNA sequences are largely based on regular expression search; in the Chomsky hierarchy the corresponding *regular* languages (RLs) are at the lowest level of expressive power. Yet, we have shown that such biologically important features as inverted repeats (and the corresponding secondary structures, e.g. stem-and-loop formations) belong to the class of *context-free* (CF) languages and not RLs. In fact, secondary structure in its full generality is *non-linear*, *non-deterministic*, and *inherently ambiguous*—all formally-defined language-theoretic properties which have significant consequences for any algorithmic approach to recognition. Generalized CF grammars have now been written for so-called "orthodox" secondary structure of nucleic acids, as well as specific grammars to recognize important instances of such structure, such as tRNA genes, in primary sequence data.

Other features of the language of DNA suggest that even CF grammars may not suffice—for instance, tandem repeats formally belong to the non-CF *copy languages*, and newly-discovered non-orthodox secondary structures called *pseudoknots* are not CF. The secondary and tertiary structure of proteins and the resulting interactions between residues suggest that genes themselves are also CF or greater, as does the nondeterminism of gene expression. A class of languages lying between CF and contextsensitive, called *indexed* languages (ILs), appears to adequately handle all the phenomena encountered thus far in biological sequences.

Interesting results have also been achieved concerning the question of the effects of genomic rearrangement on the linguistic complexity of any underlying language. It can be shown, for example, that the CF languages are not closed under the operations of duplication, inversion, or transposition—that is to say, when such evolutionary operations are applied to strings contained in a CF language, there is no guarantee that the resulting language will still be CF. Thus evolution by its nature may provide pressure toward increasing linguistic complexity. Another such pressure may arise from *superposition* of multiple levels of information, e.g. signals for successive steps in gene expression, since CF languages are also not closed under intersection.

We have used the Definite Clause Grammar (DCG) formalism associated with logic programming as a basis for developing tools for the practical analysis of sequence information on a large scale. In addition to implementing a formalism for *string variables* that appears to specify the required elements of ILs for sequence repeats, we have incorporated a number of other domain-specific features to allow for powerful pattern-matching search; these include a variant of chart parsing, imperfect matching, and special control operators for the logic-based search paradigm. We have been applying the resulting system to gene-finding applications, and have had some success in, for example, discovering all five genes in the 73kb human beta-globin cluster using a general gene grammar. We are also investigating fast software/hardware systems for *correlation* of DNA sequences, which, when integrated with our string variable grammar parsers, may allow for a large speedup in parsing such costly features.

Searls, D. B. (1988) "Representing Genetic Information with Formal Grammars" *Proc. of the 1988 National Conf. of the American Association for Artificial Intelligence*, AAAI/Morgan Kaufman, 7, 386-391.

Searls, D. B. (1989) "Investigating the Linguistics of DNA with Definite Clause Grammars" in *Logic Programming: Proc. of the N. American Conf.* (E. Lusk and R. Overbeek, eds.), MIT Press, 1, 189-208.

Searls, D. B. and Liebowitz, S. A. (1990) "Logic Grammars as a Vehicle for Syntactic Pattern Recognition" *Proc. of the Workshop on Syntactic and Structural Pattern Recognition*, IAPR, 402-422.

Cheever, E.A., Overton, G.C., and Searls, D.B. (1991) "Fast Fourier Transform-Based Correlation of DNA Sequences using Complex Plane Encoding" *Computer Applications in the Biosciences*, in press.

Searls, D. B. and Noordewier, M. O. (1991) "Pattern-Matching Search of DNA Sequences using Logic Grammars" *Proc. of the Annual Conf. on Artificial Intelligence Applications*, IEEE, 7, in press.

Searls, D. B. (1991) "The Computational Linguistics of Biological Sequences" to appear.

P45

Applying Machine Learning Techniques to DNA Sequence Analysis

Jude W. Shavlik and Michiel O. Noordewier*
University of Wisconsin, Madison, WI 53706
*Rutgers University, New Brunswick, NJ 08903

We describe a method for recognizing DNA sequences that makes use of both symbolic and neural network (connectionist) approaches to artificial intelligence.[1] The symbolic portion of our KBANN system involves applying rules that provide a roughly-correct method for sequence recognition. We then map these rules into a neural network and provide samples of sequence motifs to this network. Using these samples, the neural network's learning algorithm adjusts the rules so that they classify DNA sequences more accurately.

In an initial experiment, we have employed this method to study bacterial promoters. Rules were derived from the consensus sequences published by O'Neill.[2] Examples of promoters were taken from a compilation by Harley and Reynolds,[3] and examples of non-promoters from a DNA fragment which does not bind RNA polymerase. Our KBANN algorithm produced a network that classified *E. coli* promoters with significantly greater accuracy than published methods.

We are currently extending our studies in the following ways: 1. recognition of other DNA signal sequences such its eukaryotic splice junctions; 2. the combination of trained neural networks for the purpose of whole-gene recognition in uncharacterized sequence; and 3. development of techniques for mapping the final neural network back into the English-like vocabulary of the initial rules.

Hence, by using machine learning, we aim to produce both better classifiers of DNA sequences and better (and, importantly, human-comprehensible) theories of the structure of genes.

[1] J. W. Shavlik and G. G. Towell. "An Approach to Combining Explanation-Based and Neutral Learning Algorithms," *Connection Science: The Journal of Neural Computing, Artificial Intelligence and Cognitive Research*, **1**, 1989, pp. 233-255.

[2] M. C. O'Neill, "Escherichia coli Promotors," *Journal of Biological Chemistry,* **264**, 1989, pp. 5522-5530.

[3] C. B. Harley and R. P. Reynolds, "Analysis of E. coli Promoter Sequences," *Nucleic Acids Research,* **15**, 1987, pp. 2343-2361.

P46

Single Molecule Detection

E. B. Shera, L. M. Davis,* S. A. Soper
Los Alamos National Laboratory, Los Alamos, NM 87545
*University of Tennessee Space Institute, Tullahoma, TN 37388

It is now possible to efficiently detect and count single fluorescent molecules in solution.[1] This advance in ultrasensitive detection is an essential part of our proposed method[2] of high-speed DNA sequencing. It is also expected to have a variety of applications in biological and chemical science. The technique involves detection of the burst of fluorescence photons that occurs when a dye molecule passes through a focused laser beam. The usual spectral and spatial filtering is used to reduce interference from Rayleigh scattering and stray fluorescence. Raman scattering from the solvent, which has been the major source of background in previous experiments, is greatly suppressed by pulsed-laser excitation (70 ps at 82 MHz) and time-gated single-photon counting arranged to record only those photons that are delayed with respect to the excitation pulse (i.e., fluorescence).

A digital-filtering algorithm is applied to the signal from the photon detector. The signal-processing algorithm can be computed rapidly and it is possible to identify individual molecules in real time as they pass through the detector. Since the distribution function for the number of photons emitted by a dye molecule before photobleaching is exponential, some fraction of molecules will bleach after only a few excitations and will escape detection. However, for Rhodamine-6G (R6G) in water we have achieved a detection efficiency of 85%, with a false rate of less than 0.02 s^{-1}. We have also efficiently detected single R6G molecules in ethanol; however, this is much less challenging because the greater photostability of R6G in that solvent results in a much greater signal.

A Monte Carlo simulation of the detection technique that includes realistic geometry, appropriate photophysical dye properties, diffusion, and other relevant parameters has been developed. The simulation, which gives results in excellent agreement with our empirical observations, proved essential for optimizing the technique.

Practical utilization of the technique for various applications has led us to explore several interesting areas, among them, attachment of tagging dyes to biologically significant molecules, reduction of photobleaching, hydrodynamic focusing of molecules in solution, manipulation of objects at the molecular level, and reduction of solvent impurity background.

[1]E. B. Shera et al., Chem. Phys. Letters **174**, 553 (1990).
[2]J. Jett et al., J. Biomol. Struct. Dynamics **7**, 301 (1989).

P47

Workstation for Automated Recovery of Unstained DNA Fragments from Gels

Jeffrey M. Stiegman and Angela M. Corona
BioPhotonics Corporation, Ann Arbor, MI 48106

Many applications of electrophoresis involve separation of DNA restriction enzyme fragments not only for analysis but also for sample isolation and purification. Recovering DNA from electrophoretic gels and purifying it is an initial step in such tasks as developing cloning libraries, DNA sequencing, or sample purification after amplification using the Polymerase Chain Reaction (PCR). Present procedures for extracting DNA from a gel medium are tedious, produce inconsistent results, and consume many hours of a skilled technical staff. Such manually intensive procedures are clearly not suitable for applications requiring isolations of hundreds of DNA fragments per day.

In addition, present techniques for viewing DNA restriction fragments typically require staining with ethidium bromide dye. Ethidium bromide is an inhibitor of many of the enzymes that are commonly used in subsequent cloning steps. Removing such stains from DNA samples require additional processing steps for extraction and dialysis.

BioPhotonics has proposed to address these issues by developing a robotic workstation to provide: 1) direct visualization of electrophoretic patterns of unstained DNA using electric birefringence as a detection scheme; 2) machine vision technology to locate and excise DNA bands directly from the gel slab; and 3) automated extraction of DNA fragments from the gel matrix using a special electroelution system designed for simultaneous elution of 100-150 fragments.

To date, we have demonstrated the use of electric birefringence for recording two dimensional gel images of unstained DNA restriction fragments down to 4 Kb in length and at concentrations of 25 ng/µl. A vacuum aspirator has also been demonstrated which provides clean and precise excision of DNA bands directly from a gel without damage to the fragment and with minimal cross contamination between sampling. We have proposed to extend these developments in a prototype instrument for semi-automated recovery of unstained DNA from electrophoretic gels. Improvements have been identified which should enable recovery of unstained DNA fragments down to 100 basepairs or less and at concentration approaching 10 ng/µl with little or no cross contamination between samplings.

We expect the proposed DNA Recovery Workstation to find initial applications in those labs charged with the task of preparing the many thousands of DNA probes required to develop physical maps. Such tasks will benefit from the time savings and quality control offered by the proposed system. Future applications will be found in research and industrial sectors which will take up the task of generating large numbers of DNA probes commercially distributed for diagnostic purposes.

P48

GnomeView: A Graphical Interface to the Human Genome

David A. Thurman and Richard J. Douthart
Pacific Northwest Laboratory, Richland, WA 99352

GnomeView is a graphical user interface which displays color representations of genomic maps for review and manipulation. GnomeView's local network-model database contains chromosome maps as well as high-level, descriptive information gleaned from both GenBank® and the Genome Data Base. In response to user queries, GnomeView locates and presents maps and other information in an organized and easily used manner. Displays can be obtained for any combination of one or more chromosomes, including the entire human karyotype. GenBank features tables that can be displayed as color-coded objects mapped to the sequence representation. All maps are displayed in windows offering full zoom and pan control.

GnomeView is not a database repository. Currently, information has been extracted from existing databases, although in the future we hope to directly access the on-line databases for GenBank and GDB.

User queries can be formed to search for information in the following four ways: Name, Accession Number, Descriptive Keywords, Attributes. GnomeView returns information in the form of high-level information lists, sequence maps, or histogram representations showing the relative frequency of objects on a chromosome. In addition, the database contains links between corresponding GDB and GenBank entries allowing the user to identify and display GenBank sequences associated with a particular GDB locus.

P49

Genome Robotics and Laboratory Automation at LBL

D. C. Uber, C. M. Fockler, J. M. Jaklevic, E. H. Theil, C. R. Cantor, and C. L. Smith
Human Genome Center and Engineering Division, Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720

We describe initial efforts at developing procedures for arraying, duplication and storage of clone libraries using a commercial robot. These procedures include plate replication and preparation for storage and the production of high density filter arrays. A 96 pin tool, a sterilizer station, and an imaging station have been developed to supplement the robot's existing capabilities. Colony picking is achieved by combining image analysis with robotics. Details of operation of the system will be discussed.

Based on experience gained with these procedures, we conclude that the present approach is limited by the robot speed and methods of preparing hybridization arrays. More effective automation will require new methods of filter preparation and specialized automated workstations for the manipulation of clones and colonies. In turn, software to keep track of libraries and hybridization experiments will be required to fully exploit these new capabilities.

P50

An Integrated Intelligent System for DNA Sequence Pattern Analysis and Interpretation

E. Uberbacher, R. Einstein,* X. Guan,* R. Hand,* R. Mann, and R. Mural
Biology Division and Engineering Physics and Mathematics Division, Oak Ridge National Laboratory *Oak Ridge-University of Tennessee Graduate School of Biomedical Sciences, Oak Ridge, TN 37831

As the amount of known DNA sequence increases so does the need for computer-based tools for finding biologically relevant features in DNA sequences. We are building a hybrid neural network and rule-based inference system designed to examine and characterize regions of anonymous DNA sequence with minimal human intervention. The feasibility of automated sequence interpretation is demonstrated by the success of a new approach to feature identification using a multiple sensor-neural network formalism. This has proven to be more powerful and accurate than traditional pattern recognition methods. As an example we present a module designed to localize coding regions (coding exons) which correctly recognize 90% of coding exons of 100 or more bases with very little noise. Such individual sequence feature recognition modules represent powerful tools which can be used in a stand-alone manner before completion of the overall system. We are developing and testing other modules which will recognize features like splice-junctions, signals in the 5′ and 3′ portions of genes, etc.

Additionally we are constructing a rule-based expert system to integrate these features in to a hypothetical gene structure. In addition to biological constraints, the rule base will use scoring statistics (derived from the individual features) to determine reasonable connectivity models and select among them. Learning modifiable parameters will enable the expert system to optimize its own construction of appropriate connectivities between features. Once a tentative coding message has been extracted, comparison to sequence databases will be made automatically using highly parallel methods.

This research was supported by the Office of Health and Environmental research, United

States Department of Energy, under contract DE-AC05-84OR21400 with the Martin Marietta Energy Systems, Inc. We would particularly like to thank the Exploratory Studies Program of the Oak Ridge National Laboratory for its support.

P51

Detection of DNA Fragments in Electrophoresis Without Fluorescent or Radioactive Tags

Edward S. Yeung
Ames Laboratory and Department of Chemistry, Iowa State University, Ames, IA 50011

Our work has led to two novel approaches for the detection of native DNA fragments as they are separated by electrophoresis. The fact that fluorescent or radioactive tags are not needed means that DNA mapping procedures are greatly facilitated.

It is known that the native nucleotides absorb light around 260 nm. The application of absorption detection to the slab gel separation of DNA fragments has not been very successful because of the poor sensitivity. We developed an imaging system based on a low-noise charge-coupled device (CCD) camera. The transmitted UV light from a mercury lamp is converted to visible light by a fluorescent screen. Inhomogeneities in the gel and in the light source are corrected for by a flat-fielding computer algorithm. With this system, we were able to detect around 5 ng of DNA per band in a slab gel with exposure times of only a few seconds. We demonstrate that time-lapse sequences of images can be obtained so that separation conditions can be optimized in real time. We also show that native DNA moves faster down the gel than ethidium bromide-stained DNA, which underscores the importance of detection without chemical derivatization. A unique feature in absorption detection is that the integrated signal over each band is a good measure of the amount of DNA, since absorption coefficients for the nucleotides are well known. For fragments derived from digestion by enzymes, the relative signals therefore provide an independent estimate of the relative molecular weights for each band. This is valuable information to confirm band assignments in electrophoresis.

It has been shown that the separation of DNA fragments in capillary tubes provides better resolution and substantially shortens the run time required compared to slab gel electrophoresis. The problem has been detection of the separated components because of the low concentrations and the small amounts involved. We have successfully implemented indirect fluorescence detection in capillary electrophoresis for DNA fragments. Detection is based on charge displacement between the nucleotide and an anionic fluorophore in the buffer solution, resulting in negative fluorescence peaks. To simplify the process and to increase reliability, we used a linear polymer, hydroxypropylmethyl cellulose, in the running buffer rather than a gel inside the capillary for size discrimination. Using a 50-μm capillary tube, we were able to completely separate the fragments from a Hind III digest of λDNA in 8 minutes. Detection is possible at the pg per band range of material injected. In fact, the smaller fragments (< 1 kb) are easily visualized by indirect fluorescence, whereas they are difficult to stain with ethidium bromide in conventional detection schemes. This development opens up the possibility for rapid isolation of small quantities of DNA fragments in micro-sequencing applications.

P52

How Will Logic Programming Benefit Genome Analysis?

Kaoru Yoshida,[1] Cassandra L. Smith,[1,2] Charles R. Cantor,[1,2] and Ross Overbeek[3]
[1]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, and
[2]Human Genome Center, Lawrence Berkeley Laboratory, Berkeley, CA, 94720
[3]Argonne National Laboratory, Argonne, IL 60439

The fact that only four bases A, T, C and G can represent all the information on current life and on its history is fascinating from the viewpoint of computer science. Genome analysis is potentially a large application area for symbolic computation, since it requires advanced computer technology. As biological experimental methodology develops, more gene information is accumulated and analyzed. To proceed efficiently in the ever accelerating climate of current biological research, strong support and feedback from computer-aided analysis is mandatory. Molecular biology data is highly fluid; results must be replaced, modified or refined day by day. To keep up with this progress, it is necessary for computer analysis tools and systems to be simple and flexible.

Here we introduce logic programming as a powerful vehicle for symbolic computation: we describe (1) what it is, and (2) how beneficial it is to genome analysis. Logic programming languages provide high level data abstraction and control abstraction which allows programmers to avoid details of implementation like memory allocation and execution control that must be faced in most other programming languages. ICOT (Institute for New Generation Computer Technology, Japan) has developed a sequential inference machine, which is a personal workstation especially designed to run logic programs efficiently; ICOT has also developed the accompanying software systems which ought to be helpful for further development of genome analysis.

Many steps in biological genome analysis involve the handling of different kinds of maps. It is desirable to have a good tool with will navigate over these maps in a simple manner. Currently we are prototyping a mapping database of chromosome 21 in a logic programming language, which will let us pass from one map to another with a specific probe and let us narrow down or zoom up into each known region on a single map. We show that the internal database facility of logic programming, in which data is declared as facts, is effective for prototyping a database.

DNA Sequencing Methodology

P53

Improved Derivatives of the Transposon Gamma Delta For Amplification and Sequencing of Cloned DNAs

C. M. Berg and L. D. Strausbaugh
Molecular and Cellular Biology Department, University of Connecticut, Storrs, CT 06269-2131

The wild type transposon gamma delta (Tn1000) is a 6 kb member of the Tn3 transposable element family that has been used in our laboratories for providing mobile primer binding sites for both random (Liu et al., 1987, Nucleic Acids Res. 15:9461-9469) and non-random (Strausbaugh et al., 1990, P.N.A.S. 87:6213-17) bidirectional DNA sequence acquisition. To improve the utility of gamma delta for sequencing and PCR applications, we have generated a small, 1.8 kb derivative, mini-gamma delta, that contains a 38 bp repeat of the delta end, the 200 bp gamma delta resolution (res) site needed for efficient resolution of the transpositional cointegrates, and a 1.5 kb fragment containing the Tn5 kan gene. Mini-gamma delta also contains a number of restriction enzyme recognition sites to facilitate probe mapping for ordered DNA sequence acquisition using a single set of primers. The insertional specificity of mini-gamma delta has been examined in several cloned fragments and this element retains the random insertional properties of its wild type parent. Mini-gamma delta has been used to provide mobile primer binding sites for PCR amplification. In addition, "multiplex" mini-gamma delta derivatives have been constructed that have different oligonucleotide tags such that mixtures of DNA may be used in probe-mapping, PCR amplification, and direct dideoxy sequencing strategies.

P54

Separation of Large DNAs by Crossed Oscillating Electric and Magnetic Fields

S. B. Dev, J. Mear, Z.-Q. Xia, and G. A. Hofmann
BTX, Inc., San Diego, CA 92109

Efforts to separate large DNAs by application of Lorentz force, generated by crossed oscillating electric and magnetic fields (COEMF), have continued. We have shown that, independently, DNA of larger fragment length has higher mobility than the smaller one in plain buffer or very low strength gel, as predicted from the basic physics. This also indicates very strongly the possibility of separating mixture of DNAs by this method. Samples withdrawn from the chamber after crossfield runs, followed by pulsed field get electrophoresis show, however, that the DNAs co-migrate and there is no apparent separation. There can be several reasons for this and we are systematically exploring each of these factors. To minimize further the possibility of DNA chains overlapping, the concentration of the sample has been reduced to tenths of μg/ml.

A technique has been developed to transfer DNA directly onto nylon membrane from the solution for subsequent analysis by Southern blotting. We have also been able to obtain up to 500 kb DNA in solution, without shear breakdown, by ligation of λ DNA, thus increasing the net Lorentz force on the molecule. Use of very low concentration agarose gel and synthetic polymer have been found to reduce considerably electromagnetic convection arising out of inhomogeneity in either the electric and/or magnetic fields. Crossfield runs are also being made with Na Alginate, followed by in-situ gelation in the chamber with Calcium Chloride solution, to "freeze" the DNA bands. This may avoid band-mixing which inevitably hinders separation. We are also extending COEMF technique for very rapid concentration of DNA from dilute solution that may be useful for manipulation of high molecular weight DNA, such as YAC cloning.

Research supported by DOE SBIR Grant No. DE-AC03-88ER80655.

P55

DNA Sequencing by Hybridization: First 100 Bases Read by a Non Gel-Based Method and a Concept of Partial Sequencing Formulated

Radoje Drmanac,* Zaklina Strezoska, Tatjana Paunesku, Ivan Labat, Snezana Drmanac,
Danica Radosavljevic, and Radomir Crkvenjakov*
Institute of Molecular Genetics and Genetic Engineering, 11000 Belgrade, Yugoslavia
*Biological and Medical Research, Argonne National Laboratory, Argonne, IL 60439-4833

Determination of the sequences of human and other complex genomes requires much faster and less expensive sequencing processes compared to the methods in use today. Sequencing by hybridization (SBH) (Drmanac et al., Genomics, 4:114-128, 1989) is potentially one of such processes. We obtained hybridization data sufficient to accurately reread 100 bps of known sequence. The test and one of 5 control DNAs spotted on nylon filters were hybridized with 90 octamer and 15 nonamer probes in low temperature conditions (Drmanac et al., DNA and Cell Biol. 9:527-534, 1990). The 93 consecutive overlapping probes are derived from a 100 bp segment of test DNA and the remaining 12 probes are generated by incorporation of a non-complementary base at one of the ends of basic probes. These 12 probes have a full match target in the control DNAs. A stronger signal in DNA containing full match target compared to DNA with only end-mismatched target(s) is obtained with all 105 probes. In 3 cases (2.8%) the difference of signals is not significant (less than 2-fold) due to inefficient hybridization and consequently higher influence of background. The hybridization pattern obtained enabled us to successfully resequence the 100 bp of test DNA applying a developed algorithm (Drmanac et al., J. Mol. Struc. Dynam., in press.) which tolerates the error rate much higher than observed in the experiment. With this result, the technological components for a large scale DNA sequencing using SBH method are in place.

Partial sequencing comprises determination of a small part of constituent oligonucleotide sequences of the predetermined length within appropriate DNA fragments. This information obtained on genomic DNA or cDNA libraries has potential to reveal significantly similar (>70%) or identical sequences on the basis of comparisons of obtained oligonucleotide lists among clones. Theoretical analysis and preliminary computer simulation indicate that 500 to 3000 6-8-mer probes hybridized to 1-10 kb densely overlapped clones will generate sufficient information for this analysis. Thus partial sequencing requires 30-fold more probes and clones than mapping and 30-fold less than complete sequencing. The potential applications are in locating and counting genes and other functional sequences for which a prototype exists or finding and characterization of new ones in human and other complex genomes, i.e. in establishment of a structural genome map and a gene inventory, and in analysis of time-space patterns of gene expression by a parallel fingerprinting of over 100,000 clones per cDNA library. The main components of the immediately applicable partial sequencing procedure have been developed including the efficient growth and storage of M 13 phage in microtiter plates, spotting a sufficient amount of phage on a filter and the discriminative hybridization of spotted DNA with probes as short as hexamers. The projected rate is to collect 10 data bits per day (10 probes hybridized with million clones).

P56

Investigation of the Utility of X-ray Diffraction in DNA Sequence Analysis

J. W. Gray, J. Trebes,* D. Peters, U. Weier, D. Pinkel, T. Yorkey,* J. Brase,* D. Birdsall, and R. Rill**
Biomedical Sciences Division, *Laser Program, Lawrence Livermore National Laboratory, Livermore, CA 94550
**Department of Chemistry, Florida State University, Tallahassee, FL 32306

We report theoretical and experimental studies of the feasibility of using X-ray diffraction for rapid DNA sequence analysis. In this approach, the DNA sequence to be analyzed is amplified to ~10^{12} copies and divided into 4 fractions. In each fraction, one type of base is labeled with a heavy metal such as I or Pt that scatters X-rays efficiently. Thus, the adenines are labeled in one fraction, the guanines in the second, the thymines in the third and the cytosines in the fourth. The DNA sequence information is determined by measuring the distances between the labels in the various fractions. This is accomplished by aligning the molecules in each fraction so that they are straight and approximately parallel and recording an X-ray diffraction pattern for each. The distances between the heavy metal labels are determined by Fourier analysis of the recorded scattering patterns.

Theoretical studies employing an approximate model for the scattering of partially coherent X-rays from an oriented DNA fiber suggest that reconstruction of the positions of the heavy metal labels is possible. Experimental studies have involved the production and analysis of X-ray diffraction patterns for oriented arrays of short, heavy metal labeled oligonucleotides. DNA molecule orientation has been accomplished both by pulling DNA fibers and by preparing liquid crystalline DNA in the presence of a magnetic field. The theoretical and experimental results will be reported.

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under contract number W-7405-ENG-48.

P57

High Speed Sequencing of DNA: Preparation of Fluorescently Labelled Single DNA Molecules

Carol A. Harger, Mark L. Hammond,* John C. Martin,* Babetta L. Marrone,* and Frederic R. Fairfield
Theoretical Biology and Biophysics, *Cellular and Molecular Biology, and Center for Human Genome Studies; Los Alamos National Laboratory; Los Alamos, NM 87545

To sequence single DNA molecules by detecting the fluorescence from individual modified deoxynucleotides, we must incorporate fluorescently modified nucleotides into every position of a newly synthesized DNA strand. Sequencing of these modified DNAs by exonucleolytic degradation requires selection, attachment, and manipulation of single fluorescently modified DNA molecules.

1) DNA replication using modified nucleotides.

As analogs for fluorescently modified nucleotides, we have incorporated biotin-7-dATP and biotin-11-dUTP into M13 DNA. Recently dATP and dUTP modified with the fluorescent labels rhodamine and fluorescein attached via non-rigid linker arms have become available. Under standard incubation conditions, the incorporation of these nucleotides was not detectable. In some cases, the rate of incorporation of unmodified nucleotides was inhibited by the fluorescent nucleotides. We are developing an assay that characterizes the inhibitory effects and will allow us to determine the roles of solution conditions, DNA polymerases, and nucleotide structure on the incorporation of fluorescently labelled nucleotides into soluble double-stranded DNA. We anticipate that further modifications to the fluorescent nucleotides, such as the addition of rigid linker arms, different fluorochromes, or alternative linker arm/nucleotide attachment sites, will more readily allow DNA replication.

2) DNA selection and manipulation.

For the selection, attachment and manipulation of single DNA molecules, we are using bacteriophage lambda double-stranded DNA to simulate the size of our eventual target DNA and ethidium bromide to mimic the fluorescent tag. Using a cooled CCD camera coupled to a fluorescence microscope, we are able to see small fluorescent objects that have the mobility, size, and sensitivity to DNase expected of individual lambda DNA molecules. To manipulate individual molecules of DNA efficiently, some type of solid support is necessary. We have bound individual DNA molecules to glass beads and started to manipulate these beads. Since supported DNA must be digested one nucleotide at a time, we are investigating whether the support alters exonuclease digestion of the DNA.

P58

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

Bacteriophage T7 encodes its own DNA polymerase, the product of gene 5. The 80 kDa gene 5 protein has low processivity, dissociating from a primer-template after incorporating only a few nucleotides. Thioredoxin, a 12 kDa protein encoded by the *Escherichia coli trxA* gene, binds tightly to the gene 5 protein in a 1:1 stoichiometry and confers a high degree of processivity by increasing the affinity of gene 5 protein for a primer-template.

In addition to its DNA polymerase activity, the gene 5 protein has a 3′ to 5′ exonuclease activity. We have inactivated specifically this exonuclease activity both chemically, by a localized oxidation reaction, and genetically, by site-specific mutagenesis. The T7 DNA polymerase-thioredoxin complex which has a reduced exonuclease activity is useful for DNA sequencing by the dideoxy chain termination method. Its usefulness is enhanced by the use of $Mn2^{+}$ instead of $Mg2^{+}$ for catalysis, a substitution that eliminates discrimination against dideoxynucleoside triphosphates and generates bands of uniform intensity.

We are studying the molecular events that result in thioredoxin conferring high processivity to the polymerase reaction. We have isolated mutant thioredoxins and examined their ability to interact with wild-type gene 5 protein *in vitro*. A thioredoxin mutant which substitutes both active site cysteines for serine residues can restore nearly full polymerase activity; however, it can only do so at a 100-fold higher concentration than wild-type thioredoxin. Another thioredoxin mutant (gly-92 to asp-92) does not bind to gene 5 protein, even at extremely high concentrations.

We are also investigating the interaction of gene 5 protein and thioredoxin by genetic methods. Previously, *E. coli* thioredoxin mutants have been characterized that are unable to support the growth of wild-type T7 phage. We have used one of these thioredoxin mutants (gly-74 to asp-74) to select for T7 revertants, and we have characterized the nature of the revertant mutations in the T7 phage. The gly-74 to asp-74 mutation in the thioredoxin gene is suppressed by a replacement of glu-319 in the gene 5 protein to either a valine or lysine residue. The evidence strongly suggests that these two sites represent a contact point between thioredoxin and gene 5 protein. We have also identified other mutations within gene 5 that can suppress the same asp-74 thioredoxin mutation. We are currently identifying the location of these additional revertant mutations and are purifying the mutant proteins to investigate their biochemical properties.

P59

YAC Fingerprinting Using Cosmid Arrays: A New Strategy for Rapid Physical Mapping of Human Chromosomes

Kathy A. Lewis and Glen A. Evans
Molecular Genetics Laboratory and Center for Human Genome Research, The Salk Institute for Biological Studies, La Jolla, CA 92138

We are currently constructing a physical map of human chromosome 11 using a directed strategy including the preparation of PCR-based primers and isolation of yeast artificial chromosomes isolated from a total human genome library. While this approach is effective, it is time consuming and to expedite the mapping project, we have begun development of a separate and potentially more powerful strategy based on chromosome-specific cosmid and YAC libraries. Using this approach, chromosome-specific YAC clones are isolated from a somatic cell hybrid library and characterized by hybridization to high density arrays of chromosome-specific cosmids. We have utilized two arrayed cosmid libraries: a 1000 member 11q12-11qter library and a 16,000 member chromosome-11 specific cosmid library representing the entire chromosome.

This strategy involves a number of simple steps: 1) YAC clones are isolated from a somatic cell hybrid library at random for "fingerprinting by hybridization," 2) total yeast DNA labeled to high specific activity by random hexamer labeling with ^{32}P-nucleotides, 3) repetitive sequences are eliminated by hybridization to human repetitive sequences using a phenol emulsion hybridization procedure, 4) hybridization is carried out to a replica of the cosmid array on a nylon-backed filter, 5) the grid coordinates of cross-hybridizing cosmids are scored and entered in a computerized database. Cosmids detected by this procedure are those which represent the same segment of chromosomal DNA as the YAC DNA used as a probe. If two YAC clones detect cosmids at the same grid coordinate, the YAC clones overlap. This approach has several advantages over other YAC-based strategies for mapping including 1) detection of YAC overlaps with a high theta parameter (minimum detectable overlap), 2) elimination of false links generated by ligation artifacts frequently found in currently available YAC libraries, 3) the effective and rapid "subcloning" of YAC clones into cosmids for further analysis, 4) easy integration with other directed mapping strategies, and 5) in theory, this strategy can be multiplexed to reduce the amount of work in the analysis. This approach represents a second parallel strategy to the primary directed mapping strategy being employed for the analysis of human chromosome 11.

P60

The Development of Chemiluminescence-Based DNA Sequencing Kits

Christopher Martin and Irena Bronstein
Tropix, Inc., Bedford, MA 01730

A non-isotopic procedure for sequencing DNA by the Sanger dideoxy method has been developed. This procedure involves the use of the chemiluminescent substrate, disodium 3-(4-methoxyspiro[1,2-dioxetane-3,2'-tricyclo[3.3.1.1^{3,7}]decan]-4-yl)phenylphosphate, AMPPD® Standard DNA sequencing reactions were initiated with biotinylated oligonucleotide primers. Several DNA and RNA polymerases, including DNA polymerase I Klenow fragment, reverse transcriptase, Sequenase®, and Bacillius stearothermophilus DNA polymerase I, have been tested for use in DNA sequencing reactions with biotinylated primers. The best results were achieved with the Sequenase® and B. stearothermophilus enzymes. The resulting DNA fragments were separated by PAGE and transferred from polyacrylamide gels to nylon membranes by either electroblotting or capillary blotting. Biotinylated polynucleotides were detected after incubation of the membrane with a solution containing a streptavidin-alkaline phosphatase conjugate, followed by a solution containing the chemiluminescent substrate AMPPD.

Dephosphorylation of AMPPD by the alkaline phosphatase leads to production of the metastable anion AMP-D which further fragments to form an emitting excited state of methyl meta-oxybenzoate anion. DNA sequence data was detected by exposing the membrane to standard x-ray film for as little as one minute. Short exposure times make it possible to sequence a DNA clone in one 8-9 hour period. Our results demonstrate that DNA sequencing with chemiluminescence is a safe and efficient alternative to procedures utilizing radioactive isotopes.

Work performed under auspices of the U.S. Department of Energy, contact No. DE-AC05-89ER80752.

P61

High-Sensitivity Fluorescence Detection of DNA

Richard A. Mathies and Alexander N. Glazer
Department of Chemistry and Division of Biochemistry and Molecular Biology, University of California, Berkeley, CA 94720

Our research progress during the past year is summarized in the four publications cited below:

First, a laser-excited confocal fluorescence gel scanner has been developed and applied to the detection of fluorescent labeled DNA. An argon ion laser is focused in the gel with a high-numerical aperture microscope objective. The laser-excited fluorescence is gathered by the objective and focused on a confocal spatial filter followed by a spectral filter and photodetector. The gel is placed on a computer-controlled scan stage, and the scanned image of the gel fluorescence is stored and analyzed. A detailed technical description of this scanner as well as examples of its application to the detection of DNA separated on sequencing gels, agarose mapping gels and pulsed field gels has been presented (1).

Second, this scanner has been used to develop new DNA staining methods. We have found that ethidium homodimer binds to double-stranded DNA very tightly even under typical electrophoresis conditions. Therefore, DNA can be prestained with ethidium homodimer, electrophoresed, and then detected on agarose gels with picogram sensitivity (2).

Third, our method for optimizing the laser power and illumination time to achieve the best fluorescence detection limits has been published (3).

Finally, in more recent work we have developed staining methods which make it feasible to perform two-color detection using the laser scanner (4). The detection system for the scanner has now been modified to permit the *simultaneous* two-color detection of DNA samples labeled with ethidium homodimer and thiazole orange. Examples of two-color detection will be presented.

(1) M.A. Quesada, H.S. Rye, J.C. Gingrich, A.N. Glazer and R.A. Mathies, "High-sensitivity DNA detection with a laser-excited confocal gel scanner," Biotechniques, in press.

(2) A.N. Glazer, K. Peck and R.A. Mathies, "A stable double-stranded DNA-ethidium homodimer complex: Application to picogram fluorescence detection of DNA in agarose gels," Proc. National Academy of Science USA 87, 3851-3855 (1990).

(3) R.A. Mathies, K. Peck and L. Stryer, "Optimization of high-sensitivity fluorescence detection," Anal. Chem. 62, 1786-1791 (1990).

(4) H.S. Rye, M.A. Quesada, K. Peck, R.A. Mathies and A.N. Glazer, "High-sensitivity two-color detection of double-stranded DNA with a confocal fluorescence gel scanner using ethidium homodimer and thiazole orange", Nucleic Acids Research, in press.

P62

Isolation, Characterization and Mapping of AC-Repeat Microsatellite Sequences of Human Chromosome 16

J. C. Mulley, A. D. Thompson, Y. Shen, H. A. Phillips, R. I. Richards, and G. R. Sutherland
Department of Cytogenetics and Molecular Genetics, Adelaide Children's Hospital, North Adelaide, SA, Australia

AC-repeat microsatellite sequences from human chromosome 16 are being isolated and characterized. These are being used to establish a multipoint map of the chromosome using highly polymorphic PCR based markers and as STSs for physical mapping.

Two approaches are under way. Firstly, clones which showed hybridization to synthetic poly (AC.GT) used as a probe were randomly selected from a cosmid library. This library was constructed from the human-mouse somatic cell hybrid CY18 containing chromosome 16 as the only human chromosome. Where heterozygosity is at least 0.65 the markers are typed on the CEPH reference panel for construction of a multipoint map of the entire chromosome. The aim of this approach was not only to construct a multipoint map of the entire chromosome but also to generate markers which would by chance map to physical intervals close to regions of interest. The three regions of interest are the fragile sites *FRA16A* and *FRA16B* and Batten's disease. Heterozygosity exceeded 0.65 for two of these markers and for a third marker derived from a second approach.

The second approach involved the selection of phage and cosmid clones, which showed hybridization to the synthetic poly (AC.GT) probe, and which were previously mapped to specific regions. the AC-repeats are being characterized near the two fragile sites *FRA16A* and *FRA16B*, and Batten's disease. Where heterozygosity is at least 0.3 the marker will be typed on the CEPH reference panel for the construction of detailed linkage maps around these regions. So far, one AC marker has been characterized near *FRA16A* using this second approach. In addition, of the markers characterized using the first approach, there is another AC marker near *FRA16A*, three near Batten's disease and one near *FRA16B*.

A total of 11 markers have been characterized for their repeat sequence, number of alleles, heterozygosity and physically mapped on an extended somatic cell hybrid panel. The two most highly polymorphic markers have been typed on the CEPH panel and primer sequences for these are:

16ACXE81
Forward primer: 5′ GCT TGT ATT AGT CAG CAT TCT CAA G 3′
Reverse primer: 5′ TAC AGA CCA TAG ACT TGA CAG TCT C 3′

16AC2.3
Forward primer: 5′ GGC ATG TCA GGC CAG CCA TGT TTT 3′
Reverse primer: 5′ CCT TGC ACA AAA ACA GTA GCT ATC CAC 3′

P63
Human Chromosome 16 Physical Map: Mapping of Somatic Cell Hybrids Using Multiplex PCR Deletion Analysis of Sequence Tagged Sites

R. I. Richards, K. Holman, S. Lane, G. R. Sutherland, and D. F. Callen
Department of Cytogenetics and Molecular Genetics, Adelaide Children's Hospital, North Adelaide, SA, Australia

Physical mapping of human chromosome 16 has been undertaken using somatic call hybrid DNAs as templates for polymerase chain reaction (PCR) deletion analysis of sequenced tag sites (STS). A panel of twenty-nine somatic cell hybrids were analysed, confirming and refining previous chromosome 16 breakpoint orders and distinguishing between the locations of breakpoints in new hybrids. Ten STS markers were coamplified in three multiplex reactions allowing the rapid, simultaneous deletion analysis of nine different loci. The locations of the protamine (*PRM1*), sialophorin (*SPN*), complement component receptor 3A (*CR3A*), NAD(P)H menadione oxidoreductase 1 (*NMOR1*) and calbindin (*CALB2*) genes were refined.

P64
Synthetic Endonucleases

Betsy M. Sutherland and Gary A. Epling*
Biology Department, Brookhaven National Laboratory, Upton, NY 11973
*Chemistry Department, The University of Connecticut, Storrs, CT 06269

Construction of efficient synthetic endonucleases requires knowledge of optimum strategies for labeling, determination of binding properties of the labeled molecules, and study of the cleavage properties of the cleaving moieties alone and conjugated to the protein which confers binding specificity. We have developed new quantitative non-radioactive assays which allow determination of the extent and specificity of binding of T7

RNA polymerase labeled with rose bengal (RB-RNAP). This assay can also yield levels and specificity of cleavage by such synthetic endonucleases. We determined that standard conditions for binding of T7 RNA polymerase to DNA containing promoters first, required large excesses of RNAP per promoter, and second, also allowed polymerase binding to DNA without promoters. We therefore devised new binding conditions giving both higher levels of binding and more specificity for DNA molecules containing promoters.

We have studied the effect of different levels of rose bengal labeling of polymerase on binding and cleavage. Generally, low levels of rose bengal labeling (≤ ~ 5 RB/RNAP) allow binding of the labeled polymerase at approximately the same levels as the unlabeled polymerase. At higher levels of labeling (~ 10-15 RB/RNAP), binding of the RB-RNAP to DNA containing a T7 promoter is retained, but non-specific binding increases. We have thus chosen intermediate levels of labeling (~ 5-10 RB/RNAP) of the polymerase for cleavage of promoter-containing DNA. Under optimal conditions, the RB-RNAP cleaves DNA containing a T7 promoter but not a similar DNA lacking T7 promoter sequences.

Several strategies focusing on optimization of the cleavage properties of the cleaving moiety have been developed. These include clarification of the chemical mechanism involved in the photocleavage and modification of the chemical structure of the cleaving moiety to enhance the desired photoreactivity. Mechanistic studies suggest photo-induced electron transfer rather than singlet oxygen sensitization is the key step leading to photo-nicking. Thus, new compounds under development as potential cleaving moieties modify RB in a way that will improve its electron transfer efficiency.

P65

Transposon-Mediated Nested Deletions for Sequencing Cloned DNA

Gan Wang, Jian-Min Chen, Xiaoxin Xu, Douglas E. Berg,* and Claire M. Berg
University of Connecticut, Storrs, CT 06269
*Washington University, St. Louis, MO 63110

Transposons can be useful as they provide mobile binding sites for DNA sequencing primers in intermolecular transposition and also in intramolecular transposition (where the element moves within one replicon and generates nested deletions). Tn9 (IS1) had been used in this way (see Ahmed, 1987, Meth. Enzymol. 155:177), but its transposition is too nonrandom to be generally useful. The elements that are most random in intermolecular transposition are Tn5 and γδ, and we are investigating the behavior of small engineered derivatives of these elements during intramolecular transposition to determine their usefulness for generating nested deletions. The results with mini-Tn5 (IS50) constructs have been described (Tomcsanyi et al., 1990, J. Bacteriol. 172:6348). The mini-γδ constructs used to study intramolecular transposition contain a pair of 38 bp terminal repeats of the δ end of γδ bracketing a kan gene. Two contraselectable genes (thyA$^+$ and sacB$^+$) were cloned next to the transposon. Selection for Tmp^r (thyA$^-$), or Suc^r (sacB$^-$) mutants yielded deletions that extended into, and beyond, the target gene and inversions into that gene. Deletions and inversions into the target gene were equally frequent. All Tmp^r Suc^r double mutants had deletions that extended for varying distances in the 8.9 kb target region. No significant hotspot or coldspot was detected in this plasmid, suggesting that intramolecular γδ transposition is quite random. Thus, γδ, like Tn5, is useful for generating nested deletions that place a "universal" sequencing primer binding site at many sites in large cloned DNA segments.

P66

Combinatorial Mapping of Transposable Vectors for Sequencing Large Plasmid Inserts

R. Weiss and R. Gesteland
Howard Hughes Medical Institute and Department of Human Genetics, University of Utah, Salt Lake City, UT 84112

Efficient transposable vector systems for saturating plasmids with inserts containing common priming sites and multiplex identifier tags may have advantages over current popular vector systems, such as M13. These include: a robust *in vivo* method for producing clones which eliminates *in vitro* manipulations necessary for shotgun cloning of random fragments, bi-directional sequencing from a fixed point, and viable mapping strategies for isolating minimal spanning sets of inserts from random pools. Ordered inserts provide the minimal spanning set of clones required to complete a segment while correcting for potential hot-spotting of the transposon. Transposons also provide a method for delivering replication functions and selectable markers to large-clone inserts, which allows serial-sectioning of large plasmid inserts into smaller end-to-end clones. Utilization of these unique features of transposable vectors will reduce the effort required for going from a large-insert clone to finished DNA sequence.

Currently, initial sets of gamma-delta transposons are being used to sequence plasmids containing 10 kb inserts. The transposable vectors are conjugation-based systems that use co-integrate formation to select transposition events. Each transposon consists of 38 bp. terminal inverted repeats flanked by unique 16 bp. multiplex identifier sequences, common priming sites, an internal 34 bp loxP site required for resolution of co-integrates, and central rare-cutter restriction sites used in mapping. Two distinct configurations of these elements are being investigated: a minimal transposon containing these elements in a 260 bp. vector, and an expanded version carrying accessory functions (drug resistance and replication origins). These vectors are carried on an E. coli F plasmid pCT105 and gamma-delta transposase is provided from a cloned gene in trans. Transpositions into plasmids are isolated by mating to an F$^-$ recipient while selecting for plasmid and recipient specific drug resistances. Transposition frequency into cosmids of 4 x 10^4 transpositions/ml from a 3 hr. mating are obtainable.

Independent insertions are now being mapped and used as priming sites for sequencing 10 kb plasmids. These insertions are localized to small (100 bp) intervals by restriction site mapping of the insert site from combinatorial clone pools. The combinatorial approach allows mapping on a single gel the relative positions of 250 inserts within 10 kb fragments. The minimal spanning set is then processed through conventional Sanger dideoxy-sequencing in a multiplex fashion, allowing recovery of two divergent sequence ladders containing identifier sequence tags from each insertion.

P67

Sequence Analysis of Human Chromosome 21-Specific Linking Clones

Yiwen Zhu,[1] Cassandra L. Smith,[1,2] and Charles R. Cantor[1,2]
[1]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720
[2]Human Genome Center, Lawrence Berkeley Laboratory, Berkeley, CA 94720

Sixteen human chromosome 21-specific *Not* I linking clones have been sequenced and characterized by computer analysis. These linking clones, which are *EcoR* I or *Hind* III fragments containing a *Not* I site, were subcloned into M13 through a *Not* I clamp and/or into a modified M13 vector whose polylinker region includes a *Not* I cutting site. The sequence data shows that all 16 linking clones are G+C rich. The G+C contents vary from about 60% to more than 80%. Dinucleotide contents were also analyzed.

Most of the linking clones have very similar CpG and GpC dinucleotide contents. *Hpa* II sites occur frequently in many of these linking clones. The highest frequency is about one *Hpa* II site per 30 base pairs. This means that most of the linking clones are likely to be derived from CpG islands located at the 5′ edges of genes. Several consensus sequence patterns were found in some, but not all of the linking clones. PCR primers designed from the linking clone sequences are now being used to scan a human YAC library.

P68
DNA Sample Preparation for Scanning Tunneling Microscopy

D. P. Allison, R. J. Warmack, K. Bruce Jacobson, G. M. Brown, and T. L. Ferrell
Oak Ridge National Laboratory, Oak Ridge, TN 37831

The invention of the scanning tunneling microscope in 1982 and its subsequent development has established STM as a valuable emerging technology in biological research and, more specifically for our purposes, imaging DNA. Although most biological molecules are non-conductive in bulk form, individual molecules such as proteins and nucleic acids have been imaged when placed on conductive surfaces.

The STM operates by scanning a sharp tip a few atomic diameters away from a conductive surface and measuring fluctuations in tunneling current or voltage as either a response to changes in surface topography (STM) or changes in surface differential conductivity (scanning tunneling spectroscopy). The microscope can be operated in vacuum, air, or liquid environment with optimum resolutions of 0.1 Å vertical and 2 Å horizontal. An especially attractive feature of the technology is that images of biological molecules can be acquired without subjecting the material to staining procedures and denaturing environmental conditions.

Attachment of DNA to a suitably flat conductive surface and the ability to distinguish DNA molecules from linear surface artifacts are two major problems. We have used ^{32}P radio-labeled PBS^{+} plasmid DNA to evaluate the effectiveness of both passively and electrochemically attaching DNA to both highly oriented pyrolytic graphite (HOPG) and monatomic gold surfaces. We have also evaluated some new surfaces such as tungsten carbide, titanium carbide and iron pyrite as potential substrates for DNA. Using scanning tunneling spectroscopy we have exploited differences in conductivity to discriminate between DNA molecules and linear surface artifacts found on graphite. Further fundamental works in sample preparation will be necessary to allow routine application of STM technology to the human genome effort.

Sponsored by the Office of Health and Environmental Research, U.S. Department of Energy, under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.

P69
Increased Speed in DNA Sequencing by Utilizing SIRIS or LARIS to Localize Multiple Stable Isotope-Labeled DNA Fragments

Heinrich F. Arlinghaus and Norbert Thonnard
Atom Sciences, Inc., Oak Ridge, TN 37830

Since most current methods are slow and labor intensive, there is a critical need to develop major improvements in DNA sequence determination procedures. Given a fast, sensitive and selective detection method, large numbers of stable isotopes could be used to label DNA, thereby multiplexing the separation process, and providing a new, much faster procedure for localizing DNA after electrophoresis. At the Oak Ridge National Laboratory, methods have been developed for synthesizing tin labels for oligonucleotides, with other element labels being developed. At Atom Sciences, Sputter-Initiated Resonance

Ionization Spectroscopy (SIRIS) and Laser Atomization Resonance Ionization Spectroscopy (LARIS) have been applied to detect tin-labeled DNA. In the SIRIS/LARIS technique, the abundant neutral atoms released by the sputtering process are ionized by precisely-tuned laser beams and counted in a mass spectrometer. It is possible to localize and quantify with micrometer spatial resolution ultra-trace concentrations of a selected element at or near the surface of solid samples. The exceptionally high ionization efficiency and element specificity of the RIS process is especially valuable for ultra-trace element analysis in polymeric materials where the complexity of the matrix is frequently a serious source of interference. We have compared both the SIRIS and LARIS technique to determine their characteristics to localize and quantify Sn-labeled DNA. We will present and discuss (a) data showing differences between SIRIS and LARIS response as a function of atomization parameter, substrate and analyte, (b) the detection and resolution (spatial and isotopic) of subattomole quantities of Sn-labeled DNA bands and (c) the detection of positively hybridized and unhybridized sites on a DNA sequencing matrix. Both SIRIS and LARIS have the potential of making a strong contribution to DNA sequencing, and for any other procedures that require detection and localization of DNA, or an oligonucleotide that hybridizes to DNA. Since the ion beam and the laser beam used in the two atomization processes can be focused to a few micrometers, the electrophoresis gel length necessary for sequencing may be reduced from the 50-100 cm length currently used to 10 cm, or smaller, leading to shorter electrophoresis times, smaller quantities of DNA, less background, and faster analysis. Using Cu vapor lasers (repetition rate: 6-32 kHz) it should eventually be possible to detect more than 10^7 bases per day.

This work was supported in part by the United States Department of Energy under contract No. DE-AC-05-84OR-21400 to Martin Marietta Energy Systems, Inc., under contract No. DE-AC-05-89ER80735 and DE-FG05-90ER81015 to Atom Sciences, Inc., and by the Oak Ridge National Laboratory Exploratory Studies Program.

P70

Technology Development for Large-Scale Physical Mapping

T. J. Beugelsdijk, P. A. Medvick, R. M. Hollen, and R. S. Roberts
Robotics Section, Los Alamos National Laboratory, Los Alamos, NM 87545

The human genome effort has highlighted a huge area for potential automation. Much of the work involved in preparing maps of individual human chromosomes involves highly-repetitive procedures. Our efforts in technology development have concentrated on gridding hybridization membranes from microtiter-well plates and on database development for robotic control and for initial storage of hybridization results.

A commercial system for the construction of the hybridization array for a chromosome-specific library on a nylon filter is available. However, this product requires continual human attendance and produces small-format grids. On the basis of requirements for high-density grids, we have designed our current automated gridding system with a 30-plate dispenser and restacker to permit unattended performance.

Initial startup information is required from the operator to insure information flow about the setup. The hardware includes a NUTEC gantry robot, a Zymark microtiter-plate dispenser and restacker, a Keithley control system, a Symbol Technologies bar code reader, a disposable gridding tool, and an IBM personal computer. The software, originally written in C, has been converted to C++ in the object-oriented style of the Robot Independent Programming

Language (RIPL) developed by Sandia National Laboratories to increase maintainability. An RBase database is prepared to provide location information to the robotic arm. We can now stack 30 trays at a time for unattended gridding onto one or two membranes of one to six sectors with an interleave density of 1, 4, 9 or 16 dots per well-location.

Short-term system improvement plans include changing the sample-placement tool, adding two dispensers, and developing a plate-lid holder. Initially, we intend to use a sterilizable, reusable tool, which, along with sterilization stations, provides a more consistent surface contact with the nylon membrane, contributes sterility and lower operating cost, and reduces the plastic waste incurred with the disposable tool. Two additional microtiter-plate dispensers will permit longer unattended membrane production and will allow a full 6-sector with a 16-to-1 interleave density without refilling the dispensers. A plate-lid holder will permit the use of a more specialized tool holder on the robots arm, and it will reduce the necessary robotic movement.

Long-term system improvements include incorporating a computer workstation for integrating the initial plate-picking, membrane gridding, and hybridized membrane film assessment. Information gathered will be stored in an object-oriented database for perusal prior to entry into the laboratory workbook and the human genome database.

P71

Germ Cell Chromatin Targets for High Resolution Fluorescence *In Situ* Hybridization to Detect Closely-Spaced Genes and Multiple Copies of Related DNA Sequences

B. F. Brandriff, L. A. Gordon, A. Olsen, H. Mohrenweiser, K. Tynan, B. Trask, A. Copeland, and A. V. Carrano
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

Fluorescence *in situ* hybridization (FISH) is an integral component of efforts to construct a physical map of chromosome 19 (see abstract by Trask et al.). The resolution of FISH on metaphase chromosomes is around 1 to 2 Mbp. Mapping to somatic interphase chromatin, which is less condensed than metaphase chromatin by several orders of magnitude, has been employed as a strategy to resolve closely-spaced markers. Somatic interphase FISH mapping appears most useful for mapping probes separated by ~50 kbp to at least 1 Mbp.

The work presented here tests the hypothesis that FISH to germ cell chromatin might usefully extend the degree of resolution obtainable in somatic interphase nuclei. We used as our experimental targets the interphase nuclei present after *in vitro* fusion of human sperm with eggs from the golden hamster. In the living state, these fused eggs have a diameter of 70 μm with correspondingly large interphase nuclei, referred to as pronuclei, of ~24 μm diameter. When fixed on microscope slides, pronuclear interphase chromatin appears as an open network of chromatin strands, in contrast to somatic interphase chromatin which appears as a solid disk.

By use of a set of cosmid clones surrounding and including the hamster dihydrofolate reductase gene (Ma et al. 1988, *Mol. Cell Biol.* 8:2316) and another set of cosmids mapping within 810 kbp on human Xq28 (Kenwrick and Gitschier 1989, *Am. J. Hum. Genet.* 45:873) we showed that the mean measured pronuclear distances between hybridization sites were about three times as large as those in somatic interphases for probe pairs from the same regions. Over the range of 18 kbp to 810 kbp the correlation of physical vs genomic distance was linear and highly significant. In contrast, in somatic interphase, because of the limits of resolving power of the optical microscope, measuring distances of <50 kbp is not practical.

Another distinctive feature of the pronuclear system is the resolution of closely spaced repeat elements in the human genome, as became evident in ordering and estimating distances between contigs from the carcinoembryonic antigen (CEA)-like region on chromosome 19 (see abstract by Olsen et al.) by FISH analysis. A single probe for the pregnancy-specific glycoprotein (PSG) subgroup appeared as a string of well-separated hybridization dots, whereas the same probe appeared as a cluster of dots in somatic interphase nuclei. As part of the effort to obtain closure in this region of chromosome 19, we are establishing the relative order of contigs containing the CEA and biliary glycoprotein (BGP) genes, the PSG gene family, and an additional CEA subgroup contig. Contigs were established by fingerprinting. We estimate that the genomic distance between the CEA and BGP contigs is ~500 to 600 kbp.

The ability to resolve closely-spaced repeat elements in pronuclear chromatin will also be useful for mapping cosmids located near heterochromatic regions. When FISH was used with probes for alphoid and satellite sequences on chromosome 1, we found that this heterochromatic region remained decondensed and greatly extended throughout the first cell cycle of a fused egg, in contrast to the situation in somatic cells, where heterochromatin decondenses only briefly for replication. The same phenomenon was observed with repetitive probes for chromosomes 6, 7, 9, and the Y chromosome.

Work performed under the auspices of the U.S. DOE by the Lawrence Livermore National Laboratory under Contract No. W-7405-ENG-48.

P72

Manipulation of Single DNA Molecules Under a Fluorescence Microscope

Carlos Bustamante, Yuqiu Jiang, Steven B. Smith, Laura Finzi, and Sergio Gurrieri
Institute of Molecular Biology, University of Oregon, Eugene, OR 97403

The success of the Human Genome project depends on the ability to develop faster and more powerful methodologies to carry out the mapping and sequencing of DNA fragments much larger than those accessible though traditional methods. We are currently adapting a method originally proposed and demonstrated by Yanagida (1) and collaborators in which single DNA molecules can be visualized in real-time under the fluorescence microscope. Thus, we have recently shown that this technique can be used to study the dynamics of DNA molecules in solution (2) and inside agarose gels undergoing pulsed gel electrophoresis (3).

We plan to apply and extend this technique, to develop the ability to manipulate entire mammalian chromosomes and to perform direct restriction mapping under a microscope stage. To carry out these tasks, we must first show that it is possible to single out and manipulate individual molecules of DNA under a fluorescence microscope, using a variety of external fields.

Experiments will be described in which DNA molecules are manipulated by a variety of mechanical, electrical and chemical means. These included extending DNA molecules under applied electric fields, guiding the molecules through grooves in the glass surface, digesting them with nucleases, and picking them up, then translating and releasing them by STM tips.

The results presented here represent a first step in the development of new methods of genomic analysis. Because these methods can be easily automated they have the potential to accelerate the mapping of megabase sized DNA molecules.

Bibliography

1. Matsumoto, S., Morikawa, K., & Yanagida, M. (1981) *J. Mol. Biol.* 152: 501-516.

2. Houseal, T. W., Bustamante, C., Stump, R. F., & Maestre, M. F. (1989) *Biophys. J.* 56: 507-516.

3. Gurrieri, S., Rizzarelli, E., Beach D., and Bustamante, C. (1990) *Biochemistry* 29: 3396-3401.

P73

The Use of *In Situ* Hybridization for Characterizing Somatic Cell Hybrids, Probe Mapping and Characterizing Subtle Chromosome Rearrangements

D. F. Callen, E. Baker, H. J. Eyre, and G. R. Sutherland
Department of Cytogenetics and Molecular Genetics, Adelaide Children's Hospital, North Adelaide, SA, Australia

Chromosome painting using biotinylated DNA from a chromosome 16 specific cosmid library has allowed the detailed evaluation of three chromosome abnormalities which were initially described as deletions. Two were found to be balanced translocations and one an insertion of a portion of the short arm of chromosome 16 into the long arm of chromosome 11.

The hybrid CY170 was derived from GM2346 and this was reassessed from a deletion to a t(4;16). Similarly, CY15 was derived from T85-43A and was found to be a subtle reciprocal translocation t(1;16). The cell line HAY was used to generate the hybrid CY180. The rearrangement in the human line was an ins(11;16) which could not be distinguished by classical cytogenetic procedures.

The four fragile sites of chromosome 16 (FRA16A, FRA16E, FRA16B and FRA16D) provide useful physical landmarks since their cytogenetic location can be accurately determined. Probes have been mapped with respect to these fragile sites by *in situ* hybridization using tritiated or biotinylated probes. Two probes mapping on either side of FRA16A by *in situ* hybridization are on the same 400 kb *Sal*I fragment and thus can be utilized to clone this fragile site.

Various strategies using *in situ* hybridization with total human DNA, chromosome 16 specific libraries or specific chromosomal pericentromeric probes have been used to evaluate the chromosome content of various mouse/human hybrids. Eight monochromosomal hybrids have been isolated and twelve with 1 to 4 chromosomes.

P74

Isolation of YACs Containing Human Subtelomeric Repeats by Using Dot Blot Hybridization

Jan-Fang Cheng
Human Genome Center, Lawrence Berkeley Laboratory, Berkeley, CA 94720

Human telomeres contain conserved T_2AG_3 terminal repeats and different types of subtelomeric repeat sequences. One type of subtelomeric repeats (designated as A1 repeat) is found in multiple copies adjacent to many telomeric T_2AG_3 repeats. I have also detected copies of A1 repeats located at least 50 kb away from some chromosomal ends. These results indicate that the A1 repeat sequence can be used to isolate YACs from the vicinities of many human telomeres.

Dot blot hybridization was used to screen a genomic YAC library for the A1 containing clones. Briefly, small numbers of YAC clones (i.e. 12 and 96) were pooled, and DNAs were isolated from these pools in 96-well microtiter tubes. DNAs isolated from these pools were then spotted onto nylon filters by using dot blot apparatus. For one genome coverage, we need to screen approximately 10,000 YACs with an average size of 200 kb. This number of YAC clones can almost be placed onto one filter with 96 dots and a pool size of 96 on each dot, or eight filters with 96 dots and a pool size of 12.

The method we used to screen for the A1 containing clones is equally useful in screening a genomic YAC library with a large number of probes. For example, one can combine a large number of chromosome specific probes and screen for chromosome specific YACs. Alternatively, one can produce DNA probes from somatic hybrid cells with partially deleted human chromosomes by inter-Alu PCR amplification, and screen for YACs that are located in a specific chromosome region.

P75

Construction of Partial Digest Libraries from Flow Sorted Human Chromosomes

L. L. Deaven, M. K. McCormick, C. E. Hildebrand, R. K. Moyzis, N. C. Brown, E. C. Campbell, M. L. Campbell, J. J. Fawcett, A. Martinez, L. J. Meincke, P. L. Schor, and J. L. Longmire
Center for Human Genome Studies and Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545

The National Laboratory Gene Library Project is a cooperative project between the Los Alamos and Lawrence Livermore National Laboratories. At Los Alamos, a set of complete digest libraries has been cloned into the EcoRI insertion site of Charon 21A. These libraries are available from the American Type Culture Collection, Rockville, MD. We are currently constructing sets of partial digest libraries in the cosmid vector, sCos1, and in the phage vector, Charon 40, for human chromosomes 4, 5, 6, 8, 10, 11, 13, 14, 15, 16, 17, 20, and X. Individual human chromosomes

are sorted from rodent-human cell lines until approximately 1 μg of DNA has been accumulated. The sorted chromosomes are examined for purity by in situ hybridization. DNA is extracted, partially digested with Sau3A1, dephosphorylated, and cloned into sCos1 or Charon 40. Partial digest libraries have been constructed for chromosomes 4, 5, 6, 8, 11, 13, 16, 17, and X. Purity estimates from sorted chromosomes, flow karyotype analysis, and plaque or colony hybridization indicate that most of these libraries are 90-95% pure. Additional cosmid library constructions and 5-10X arrays of libraries into microtiter plates are in progress. Libraries have also been constructed in M13 or bluescript vectors to generate STS markers for selection of chromosome specific inserts from a genomic YAC library. We have also been able to clone sorted DNA into YAC vectors and expect to be able to construct YAC libraries representing individual chromosomes.

Supported by the U.S. Department of Energy under contract W-7405-ENG-36.

P76

Double Minute Chromosomes as Multimegabase Cloning Vehicles

Peter J. Hahn, John Hozier,* and Michael J. Lane
State University of New York Health Science Center, Syracuse, NY 13210
*Florida Institute of Technology, Melbourne, FL 32901-6988

A need exists for a cloning vehicle capable of maintaining DNA segments intermediate in size between whole chromosomes and YACs or cosmids. We have developed a system for using double minute chromosomes in mouse EMT-6 cells that can "clone" mammalian chromosomal fragments in the 1 to 10 megabase range suitable for sub-chromosomal library construction. To clone DNA in this size range, we first introduce linked Neo and dihydrofolate reductase (DHFR) genes into the genome to be sub-cloned. This is followed by lethal irradiation and fusion to mouse EMT-6 cells, and selection for neo resistance. Subsequent selection for increasing levels of methotrexate (MTX) resistance leads to amplification of the introduced segment. To demonstrate the feasibility of this system, we have co-transfected Neo and DHFR genes into Chinese hamster ovary (CHO) cells, and transferred the CHO fragment containing the linked genes to mouse EMT-6 cells by radiation/fusion hybridization, and amplified the introduced chromosomal segment by selection for resistance to increasing levels of MTX. We have used pulsed-field gel electrophoresis to map approximately 700 kb surrounding the introduced Neo, DHFR genes in the CHO donor line and in five EMT-6 radiation fusion isolates that received the same CHO chromosomal segment. Four of the five radiation/fusion hybrids received segments with maps indistinguishable from the donor - the fifth presumably had a radiation breakage within the 700 kb segment we have mapped. No further rearrangements have been detected in the subsequent amplification steps. All lines contain cytogenetically observable double minute chromosomes, and the introduced genes are unstable in the absence of selection. These data suggest that this system should be ideal for radiation/reduction type sub-chromosomal libraries because chromosomal fragments not attached to the selective marker will be readily lost with time, and the small size of double minute chromosomes should facilitate isolation of the introduced DNA.

This genome fragmentation strategy should allow construction of comprehensive human genome libraries containing less than 10,000 members while retaining high redundancy. We report the methodology we are employing to construct a partial library of human chromosome 17q. The process involves infection of a somatic cell hybrid containing human 17q with the defective retrovirus pZipneo, lethal irradiation and screening of several hundred G418 resistant EMT-6 recipient isolates for the presence of human DNA. Characterization of several of the human DNA containing EMT-6 cells identified, using inter-Alu PCR and *in situ* hybridization, reveals that this strategy can be used to produce selectable human double minute chromosomes in the mouse EMT-6 cell line.

P77

Gold Cluster Labeling of RNA and DNA and High Resolution Scanning Transmission Electron Microscopy (STEM)

James F. Hainfeld, Phillip Rappa, Mathias Sprinzl,* Valsan Mandiyan,** Santa J. Tumminia,** and Miloslav Boublik**
Brookhaven National Laboratory, Upton, Long Island, NY 11973
*Laboratorium fur Biochemie, Bayreuth, Germany
**Roche Institute of Molecular Biology, Nutley, NJ 07110-1199

Yeast tRNA^{Phe} was altered to enzymatically introduce 2-thiocytidine (s^{2}C) at position 75. This was covalently coupled to an undecagold cluster which has a 0.8 nm gold core. Similarly E. coli tRNA^{Arg} with a naturally occurring s^{2}C at position 32 in the anticodon region was labeled, but required denaturation. These were visualized in the STEM and are the smallest nucleic acid labels yet found that are stable in the electron microscope giving a resolution of ~1 nm or ~3 base pairs. This is the first example of high resolution RNA labeling by electron microscope.

A second project was labeling a palindromic 12 mer with a 5′ sulfhydryl with the gold cluster. Upon renaturation one would expect a short double stranded piece of DNA separating a gold cluster at each end. In preliminary data, pairs of clusters were observed at the appropriate distance. This is the first time that a piece of DNA as small as a 12 mer could be discerned and identified directly (by these labels) in the electron microscope.

P78

Atomic Force Microscopy Studies of Uncoated T4 Bacteriophage on Si Substrates

W. Kolbe, D. F. Ogletree, and M. Salmeron
Human Genome Center and Engineering and Materials Science Divisions, Lawrence Berkeley Laboratory, Berkeley, CA 94720

A new Atomic Force Microscope (AFM) employing laser beam deflection from a microfabricated cantilever has been constructed and applied to studies of biological material. In our first test study we used T4 bacteriophage virus deposited on electronic grade silicon wafers. The viruses were prepared by diluting stock solution in pure water to various concentrations. The solution contained buffer salts of trisEDTA. Microliter droplets of the solution were deposited and allowed to dry onto the Si substrate. The AFM images were obtained in the repulsive mode both in air and in water. The images show the virus isolated as well as forming aggregates of various sizes. The external structure as well as strands of DNA streaming out of the virus could be observed. The AFM images are compared with Scanning Electron and Epifluorescence Microscope images.

P79

Scanning Molecular Exciton Microscopy: A New Approach to Gene Sequencing

Raoul Kopelman, John Langmore,* Vladimir Makarov,* and Bradford Orr**
Department of Chemistry, *Biophysics Research Division, and **Department of Physics, University of Michigan, Ann Arbor, MI 48109

Molecular Exciton Microscopy (MEM) is a new principle of imaging, based on short-range optical interactions between a specimen and a microscopic optical probe that is scanned over the surface of the specimen. The microscope depends on the well-understood mechanics of other scanned-tip microscopies such as Scanning Tunneling Microscopy (STM), but is designed to take advantage of the many well-understood optical interactions that occur between molecules separated by 0.1 - 5 nm. Previously, photon-emitting tips have been produced using glass capillaries drawn to small dimensions. Such tips are severely limited due to the difficulty in transmitting photons through sub-wavelength apertures of the capillaries. Our recent developments have overcome these intensity problems by using excitons. A small exciton-conducting organic or inorganic crystal is grown in the aperture region and illuminated with laser light. The energy, in the form of excitons, is conducted through the crystal to the tip of the micro-capillary. The energy is converted again to light, forming an extremely small, bright source of light (Lieberman et al., Science 247:59). Our interest, however, is to take advantage of non-radiative interactions between molecules on the tip of the capillary and those on the specimen. The simplest specimens will be individual molecules bound to atomically smooth substrates. As the tip is scanned along a flat surface parallel to the substrate, the optical interactions between the tip and the specimen will be detected. The resolution of the microscope will be limited by the size of the optical source (dependent upon the geometry of the donor molecule or molecules at the tip), the distance between the source and the specimen ($\geq$O.2 nm), and the range of the optical interaction. One interaction useful at low resolution would be Förster-Dexter energy transfer, which would change in the intensity and wavelength of photons emitted from both the tip and the sample. A much higher resolution interaction would be spin-orbit coupling between specific molecules bound to the tip of the capillary and heavy atoms in the sample (known in molecular optics as the "external heavy atom effect"). In principle the interaction has a range of less than ~0.5 nm and is very effective at stimulating emission of light at a new wavelength. Photomultipliers placed above or below the specimen will detect specific emissions from the tip or the specimen, by utilization of optical filters. As the tip is scanned the detected signals will be stored on a computer.

The MEM offers many possibilities for improving gene mapping and sequencing techniques. For example, if single-stranded DNA with mercury atoms on specific bases can be bound to flat surfaces, the "external heavy atom effect" could be used to localize the heavy atoms to high resolution in order to "read" specific bases along the strands. Use of fluorescent endlabels and intelligent scanning algorithms might allow the labeled bases to be mapped at high speed. These, and other optical interactions, could be exploited in order to extend the usefulness of optical interactions to near-atomic resolution in a number of potential applications to molecular and cell biology.

P80

Imaging of Biomolecules with the STM and AFM

T. Wilson, M. Murray, M. Bednarski, G. Neubauer, W. Kolbe, C. Cantor, D. F. Ogletree, and M. Salmeron
Human Genome Center and Materials Science Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720

Most efforts to image DNA are unsuccessful due to the strong forces that the tip exerts on the surface in conventional STM and AFM instruments. These forces displace the weakly bound material and prevent imaging. To solve these problems, we have addressed the following points during last year: (1) Develop a new STM instrument that operates at gap resistances or tip surface separations sufficiently large to decrease the forces exerted by the tip. Our new instrument has produced the desired results for molecules with thickness of up to 40 Angstroms. (2) Develop and apply a combined STM/AFM to simultaneously measure current and forces as the tip approaches the surface that is covered with an organic layer of well defined thickness. In this manner we investigate the electron transport properties of large, insulating biomolecules. (3) Develop fixation techniques to ensure better reproducibility and to provide ways of orienting DNA strands on the substrate for future sequencing. The fixation is performed via intermediate bifunctional molecules binding to the substrate on one end and to DNA on the other. Examples include quaternary aminesilane for ionic binding to the DNA backbone, alkylsilanes with DNA intercalators as terminal groups etc.

P81

Ultrasensitive Luminescence Detection of Lanthanide Ion Labels for DNA Sequencing and Mapping

J. Michael Ramsey, William B. Whitten, Gilbert M. Brown,* Martha L. Garrity,*
K. Bruce Jacobson,** and Richard L. Sachleben*
Analytical Chemistry Division, *Chemistry Division, and **Biology Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6142

We are developing a new approach to rapid DNA sequencing which involves a new luminescent labeling system, an ultrasensitive detection system and capillary gel electrophoresis. The lanthanide ions Eu(III), Tb(III), Sm(III), and Dy(III) will be used as labels. These ions will be attached to a primer with a derivative of the macrocyclic chelating agent 1,4,7,11-tetraazacyclododecane-1,4,7,11-tetraacetic acid (DOTA). DNA fragments generated by the Sanger technique will be separated by capillary gel electrophoresis. Detection takes place in microdroplets of effluent that are confined in an electrodynamic trap where laser excited fluorescence spectroscopy is performed. Previous experiments with this detection technique indicate the ability to detect single molecules of Rhodamine 6G. Detection of the lanthanide ions is more challenging due to their lower quantum efficiencies and absorption cross sections. The longer luminescence lifetimes of lanthanide ions does allow for the implementation of gated detection with relatively simple hardware providing significant rejection of background signals. The narrow bandwidth emission from the Ln(III) ions should allow multiple labels to be detected simultaneously with little interference. Preliminary work on the design of the luminescent tags and development of the detection system will be presented.

Research sponsored by the Exploratory Studies Program of the Oak Ridge National Laboratory and the Office of Health and Environmental Research, U.S. Department of Energy, under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.

P82

Scanning Tunneling Microscope Images of Adenine and Thymine at Atomic Resolution

W. Siekhaus, M. J. Allen,* M. Balooch, S. Subbiah,** R. J. Tench,*** and R. Balhorn*
Chemistry and Material Science, *Biomedical Sciences Division, and ***Department of Applied Sciences, Lawrence Livermore National Laboratory, Livermore, CA 94550
**Department of Cell Biology, Beckman Laboratories for Structural Biology, Stanford University School of Medicine, Stanford, CA 94305

The scanning tunneling microscope has been used to obtain images of DNA that reveal its major and minor grooves and the direction of helical coiling, but sufficient resolution has not yet been achieved to identify its bases. To determine if this technology is capable of identifying individual DNA bases, we have examined the molecular arrangements of adenine and thymine (two of the four DNA bases that comprise the genetic code) attached to the basal plane of highly oriented pyrolytic graphite. Both molecules form highly organized lattices following deposition on heated graphite. Lattice dimensions, structural periodicities, and the epitaxy of adenine and thymine molecules with respect to the basal plane of graphite have been determined. Images of these molecules at atomic resolution reveal that the aromatic regions are strongly detected in both molecules while the various side-groups are not well-resolved. These studies provide the first evidence that tunneling microscopy can be used to discriminate between purines and pyrimidines.

This work was performed under the auspices of the U.S. D.O.E. by Lawrence Livermore National Laboratory under Contract No. W-7405-ENG-48.

P83

A Multi-Angle, Multi-State CHEF Instrument with Algorithms for Optimizing Pulsed Field Parameters

S. Ferris, D. Wilner, W. Stubblebine, C. Ragsdale, S. Freeby, and P. Zoller
Bio-Rad Laboratories, Hercules, CA 94547

We have developed a new CHEF instrument, the CHEF MAPPER™ to facilitate mapping and other applications of pulsed field electrophoresis. A key feature is the ability to electronically change the angle of pulsing from 0 to 360°. Using a 106° angle, for example, we can separate the chromosomes of *S. pombe* (3-6 Mb) in about 24 hours. The instrument includes an advanced algorithm for optimizing pulsed field parameters. The user inputs the smallest and largest fragments in a sample, then the embedded algorithm computes the switch time (or ramp), switch angle, voltage and run time required for maximum band mobility. Additional parameters, such as agarose concentration, can be manipulated from an enhanced algorithm running on a PC. Information can be transferred to the CHEF MAPPER by RS232, bar code, or by manual entry. For DNAs smaller than 50 kb, the algorithm implements a field inversion mode with the capability of different forward and reverse voltages. For advanced DNA manipulations, the user may program the CHEF MAPPER to combine up to 15 vectors (angle, voltage, duration) per switch cycle.

An eight state separation was used to resolve doublet bands of *S. cerevisiae*, difficult or impossible by other methods. Data is also presented showing how a non-linear switch time ramp extends the linear range of separation of lambda ladder. Further data are presented which illustrate various aspects of the machine's performance.

The CHEF MAPPER is based on recently developed PACE technology. The CHEF MAPPER can electronically adjust electrode potentials during a run to maintain highly uniform electric fields. The unit combines the power supply, switching, and driver functions in a single module with battery back-up of memory.

P84

DNA Polymerase Specificity for Rhodamine, Fluorescein and Biotin Labeled Nucleotides

Roger S. Lasken, Alberto Haces, and Gulilat Gebeyehu
Bethesda Research Laboratories, Life Technologies, Inc., Gaithersburg, MD 20877

Eight different DNA polymerases were compared for utilization of N4-rhodamine-dCTP (Rh-dCTP) and N6-rhodamine-dATP (Rh-dATP). Rh-dCTP could be incorporated into products 100-200 nucleotides in length although rates of synthesis were about 40 times slower than for dCTP. In general, the greatest incorporation occurred for processive polymerases lacking 3′-exonuclease activity such as modified T7 and T5 polymerases (genetically altered to inactivate the exonuclease) and Taq polymerases. Only low levels of incorporation were achieved with E. coli pol III holoenzyme, the Klenow fragment, T4 and native T7 polymerases and AMV reverse transcriptase. Rh-dATP was an inefficient substrate blocking elongation by all of the polymerases.

Competition studies suggest that the Rh-dNTPs are rapidly inserted but inhibit subsequent primer elongation. Consistent with this interpretation, DNA sequencing gels show that consecutive runs of incorporation present the most severe blocks to the polymerase.

Consecutive incorporations presumably create base stacking distortions that are difficult for the polymerase to recognize or result in increased melting of the primer terminus.

The Klenow fragment and modified T7 and T5 polymerases were also compared for utilization of biotinylated and fluorescein labeled nucleotides. The nucleotides contained biotin attached through a 14 atom linker arm to the N6 position of dATP (Bio-14-dATP), the N4 position of dCTP (Bio-14-dCTP) of fluorescein attached to the N6 position of dATP (Fl-dATP). Rates of synthesis were remarkably polymerase specific. T5 polymerase incorporated the most bio-14-dATP, T7 polymerase was most active for Fl-dATP while T7, T5 and Klenow fragment were about equal for Bio-14-dCTP. The results suggest that every nucleotide analog of interest should be individually tested with many different polymerases.

In further studies, the accuracy of DNA synthesis was determined for each nucleotide analog using modified T7 polymerase. Biotinylated dATP analogs were exclusively inserted opposite template T sites. Thus, these nucleotides form "correct" base pairs in spite of the modification of the amino group. dTTP and dCTP analogs tended to form their respective purine pyrimidine base mismatches. Rh-dUTP substituted for dCTP forming Rh-U·G mismatches and Rh-, Bio-7-, and Bio-14- dCTP substituted for dTTP in mismatching with template A.

P85

A Fractal Representation Approach to Classify the Functional Regions of DNA Sequences[1]

H. A. Lim[2, 3]
Supercomputer Computations Research Institute, Florida State University, Tallahassee,
FL 32306-4052

The databases at genomic repository centers are augmented by about twofold every year, but the functional roles of some of the genomes' fragments are still not well-understood. It is proposed that a fractal representation of a set of sequences (FRS), which can reflect the local and global patterns of gene sequences, can be used to plot sequences with similar functions. By introducing a similarity measure, a number of fully sequenced genes: actins, globins, interferons and others are classified. The approach is further used to analyze the exon and intron sequences from the human genome. The analysis shows that the FRS of introns and exons are very distinct and that there are simple repeats of A_mG_n and T_rC_s, where m, n, r, s are integers, in introns. The approach may have a number of applications. For example, as new sets of different sequences with specific functions are accumulated (sequenced), it will be possible to create reliable masks for each functional set and apply these masks for the classification of new sequences. Furthermore, histograms of gene sets which are constructed using masks of the genes display the degree of homogeneity of the sets. This can be used for detecting possible unknown sequences in a given set. An advantage of the approach is that the method is easy to implement and that it does not require the traditional alignment procedure of analyzing sequences.

[1] Work supported by grants under contract numbers TRDA 116 (Florida TRDA), DE-FC05-85ER250000 (US DOE).

[2] Abstract submitted to the Human Genome Workshop, February 17-21, 1991, Santa Fe, New Mexico, USA.

[3] Work done in collaboration with S.V. Korolev of Engelhardt Institute of Molecular Biology, USSR Academy of Sciences, Vavilov Street 32, 117984 Moscow B-334, USSR and V.V. Solovyev of Institute of Cytology and Genetics, USSR Academy of Sciences, Siberian Branch, Lavrentyeva Str. 10, 630090 Novosibirsk, USSR. SVK and VVS are partially supported by the USSR Human Genome Program.

P86

An Automated Procedure for Loading DNA Sequencing Gels and Reading DNA Sequencing Films

C. Ragsdale, W. Stubblebine, A. Tumolo, R. K. Wilson,* F. Witney, and A. Zrolka
Bio-Rad Laboratories, Richmond, CA 94806
*Department of Genetics, Washington University School of Medicine, St Louis, MO 63110

Manually loading DNA sequencing gels often yields autoradiograms with a variety of artifacts. This results in an overall reduction in the amount of data generated per gel. In addition, loading an ultra-thin sequencing gel is a difficult and tedious task, which can lead to misloading when many samples are involved. Here we report on an automatic loader for DNA sequencing gels that eliminates the artifacts and errors associated with manual loading and reduces the time and labor spent by the operator. The autoloader can reproducibly pick up small volumes from either a microtiter plate or microfuge tube array and can accurately deliver the samples to the gel in the user defined configuration. The

results obtained with the autoloader were shown to be consistently as good or better than those obtained with manual loading.

We have also developed an instrument to automate reading of DNA sequencing films.

The reader utilizes a CCD linear array detector to acquire an image of the sequencing film, and an advanced signal processing algorithm to automatically make base calls and create on-screen review image. We have analyzed over 200 lane sets from a variety of sources and found that the average base calling accuracy was in excess of 98%. The size of the data set was sufficiently large, in excess of 50,000 base, to ensure that the typical artifacts associated with DNA sequencing films were encountered. The results from this study allow us to conclude that the film reader is capable of interpreting routine DNA sequencing films at an accuracy level suitable for large scale DNA sequencing projects. In addition examples of the functioning of the review and editing system used to manually resolve ambiguous regions of sequence ladders will be presented.

P87

Single Molecule Detection of Nucleotides Tagged with Fluorescent Dyes

S. A. Soper, L. M. Davis, J. H. Jett, R. A. Keller, J. C. Martin, and E. B. Shera
Los Alamos National Laboratory, Los Alamos, NM 87545

Several photophysical parameters play important roles in determining the feasibility to detect individual molecules using ultrasensitive laser-induced fluorescence. The photophysical constants that need to be investigated when selecting candidates for single molecule detection include the absorption cross section, fluorescence lifetime and the photodestruction efficiency. Molecular photostability is especially important because it sets an upper limit on the number of photons attainable per molecule. We have measured the photophysical parameters of several different fluorescent dyes and dyes attached to various analytes (i.e., nucleotides).

The measured photostability and quantum efficiency of rhodamine 6G (R6G) and adenine labeled with tetramethylrhodamine isothiocyanate (TRITC-AD) indicate that these molecules yield similar numbers of photons per molecule. The fluorescent quantum yield of TRITC-AD is nearly three times smaller than the quantum yield for R6G, but due to its increased photostability, both fluorophores yield approximately the same number of photons. The TRITC-AD fluorophore is the type of fluorophore that will be encountered in the rapid sequencing of DNA methodology recently proposed by our group (1).

In light of our success in detecting single molecules of R6G using pulsed-laser excitation with time-gated detection (2), we have extended this work to TRITC-AD. We are able to observe the photon burst from individual molecules of TRITC-AD with efficiencies >70% and low error rates. Single molecule detection results for fluorescent dyes attached to nucleotides using our pulsed-laser single molecule detection apparatus, along with the measured photophysical parameters of several different fluorophores will be presented.

1. J. H. Jett et al., J. Biomol. Struct. and Dynamics, 7, 301 (1989).

2. E. B. Shera et al., Chem. Phys. Lett., 174, 553 (1990).

P88

BISP: VLSI Solutions to Sequence Comparison Problems

Timothy Hunkapiller, Leroy Hood, Ed Chen,* and Michael Waterman**
Department of Biology, California Institute of Technology, Pasadena, CA 91125
*Jet Propulsion Laboratory, Pasadena, CA 91109
**University of Southern California, Los Angeles, CA 90089-1113

Our research has been focused on redefining the dynamic programming paradigm such that not only could these methods be implemented effectively in silicon, but also provide, in that context, the flexibility required of a robust biological tool. BISP represents a systolic implementation of a dynamic programming algorithm based on that of Smith and Waterman, optimizing its ability to define local similarities. There are three fundamental functional differences between the BISP and previous efforts: 1) BISP is optimized for the determination of any number of local similarities between pairs of sequences (although the algorithm is also capable of returning global comparison values); 2) BISP returns values that will allow for the reconstruction of the alignment; and 3) BISP is specifically designed to employ complex, user-definable similarity rules. Most significantly, BISP supports user-selectable alphabets of up to 128 characters, complete indel penalty definition and completely individually-defined similarity values between characters in the chosen character set.

Author Index

Workshop Participants

Some California 415 area codes have changed since this list was prepared.

Michelle Alegria
Lawrence Livermore National Laboratory
P. O. Box 5507
Livermore, CA 94550
415/422-4098
Fax: 415/423-3608

David P. Allison
Oak Ridge National Laboratory
P. O. Box 2008
Oak Ridge, TN 37831-6264
615/574-5823

Chris Amemiya
Human Genome Center
Lawrence Livermore National Laboratory
P. O. Box 5507, L-452
Livermore, CA 94551
415/423-7687
Fax: 415/423-3608

Norman G. Anderson
Large Scale Biology Corporation
9620 Medical Center Drive, Suite 201
Rockville, MD 20850-3300
301/424-5989
Fax: 301/762-4892

Stylianos E. Antonarakis
The Johns Hopkins University
School of Medicine
Baltimore, MD 21205
301/955-7872
Fax: 301/955-0484

Linda Ashworth
Lawrence Livermore National Laboratory
P. O. Box 5507
Livermore, CA 94550
415/422-5665
Fax: 415/423-3608
linda@snrp.llnl.gov

Charalampos Aslanidis
Human Genome Center
Lawrence Livermore National Laboratory
P. O. Box 5507, L-452
Livermore, CA 94551
415/423-8151
Fax: 415/423-3608

Raghbir S. Athwal
New Jersey Medical School
185 South Orange Avenue
Newark, NJ 07109
201/456-5215
Fax: 201/456-3644

David F. Barker
University of Utah
Department of Medical Informatics
420 Chipeta Way #180
Salt Lake City, UT 84108
801/581-5070
Fax: 801/585-3232 o-6052

Benjamin J. Barnhart
Human Genome Program
U.S. Department of Energy
OHER, ER-72, F-201 GTN
Washington, DC 20545-0001
301/353-5037
Fax: 301/353-5051
barnhart@oerv01.er.doe.gov

David Benton
NIH National Center for
Human Genome Research
Building 38A, Room 610
9000 Rockville Pike
Bethesda, MD 20892
301/496-7531
Fax: 301/480-2770
benton@bio.nlm.nih.gov

Claire M. Berg
The University of Connecticut
Box U-131
354 Mansfield Road
Storrs, CT 06269-2131
203/486-2916 office
Fax: 203/496-1936
berg@uconnvm

Douglas E. Berg
Washington University Medical School
Microbiology Department
Box 8230, 660 South Euclid Avenue
St. Louis, MO 63110-1093
314/362-2772
Fax: 314/362-1232

George Bers
Bio-Rad Laboratories
15111 San Pablo Avenue
Richmond, CA 94806

Bruce W. Birren
California Institute of Technology
Biology Division, 147-75
Pasadena, CA 91125
818/356-4504
Fax: 818/796-7066

David Botstein
Stanford University School of Medicine
Department of Genetics
Stanford, CA 94305-5120
415/723-3488
Fax: 415/723-7016

E. Morton Bradbury
Los Alamos National Laboratory
Life Sciences Division
Los Alamos, NM 87545
505/667-2690
Fax: 505/665-3024

Brigitte Brandriff
Human Genome Center
Lawrence Livermore National Laboratory
P. O. Box 5507, L-452
Livermore, CA 94551
415/423-0758
Fax: 415/423-3608

Elbert Branscomb
Human Genome Center
Lawrence Livermore National Laboratory
P. O. Box 5507, L-452
Livermore, CA 94551
415/22-5681 Fax: 415/423-3608
elbert@alu.llnl.gov

Irena Bronstein
Tropix, Inc.
47 Wiggins Avenue
Bedford, MA 01730
617/271-0045
Fax: 617/275-8581

Henry T. Brown
Los Alamos National Laboratory
Group T-10, MS K710
Los Alamos, NM 87545
505/665-2591
htb%life@lanl.gov

Christian Burks
Los Alamos National Laboratory
Group T-10, MS K710
Los Alamos, NM 87545
505/667-6683
cb%intron@lanl.gov

Carlos Bustamante
University of Oregon
Institute of Molecular Biology
1827 Fircrest Drive
Eugene, OR 97403
503/346-1537
Fax: 503/346-5891

David Callen
Adelaide Children's Hospital
Department of Cytogenetics
72 King William Road
North Adelaide S.A. 5006
618/267-7284
Fax: 618/267-7342

Charles Cantor
Human Genome Project
Lawrence Berkeley Laboratory
One Cyclotron Road, MS 2-300
Berkeley, CA 94720
415/486-6800
Fax: 415/486-5282

John Carpenter
Cray Research, Inc.
655 East Lone Oak Drive
Eagan, NM 55121
612/683-3633
Fax: 612/683-3099

Anthony Carrano
Human Genome Center
Lawrence Livermore National Laboratory
P. O. Box 5507, L-452
Livermore, CA 94551
415/422-5698
Fax: 415/423-3608
avc@sts.llnl.gov

C. Thomas Caskey
Baylor College of Medicine
Institute for Molecular Genetics
One Baylor Plaza, T-809
Houston, TX 77030
713/798-4774
Fax: 713/798-7383

Larry Chavez
Vivigen, Inc.
2000 Vivigen Way
Santa Fe, NM 87505
505/438-1111
Fax: 505/438-1101

C. H. Chen
Oak Ridge National Laboratory
Chem. Physics Sec., Photophysics
P. O. Box 2008
Oak Ridge, TN 37831-6378

Jan-Fang Cheng
Human Genome Center
Lawrence Berkeley Laboratory
One Cyclotron Road, MS 74-3110
Berkeley, CA 94720
415/486-6549
Fax: 415/486-6816

George Church
Harvard Medical School
Department of Genetics
20 Shattuck Street
Boston, MA 02115
617/732-7562
Fax: 617/732-7663
church@rascal.bwh.harvard.edu

Michael Cinkosky
Los Alamos National Laboratory
Los Alamos, NM 87545
505/665-0840
Fax: 505/665-3493
michael@genome.lanl.gov

Nikki Cooper
Los Alamos National Laboratory
Los Alamos Science, M708
Los Alamos, NM 87545
505/667-1447

L. Scott Cram
Los Alamos National Laboratory
Life Sciences Division
Los Alamos, NM 87545
505/667-2690
Fax: 505/665-3024

Radomir Crkvenjakov
Argonne National Laboratory
Biological & Medical Research Division
9700 South Cass Avenue
Argonne, IL 60439-4833

Pieter de Jong
Human Genome Center
Lawrence Livermore National Laboratory
P. O. Box 5507, L-452
Livermore, CA 94551
415/423-8145
Fax: 415/423-3608
pieter@pcr.llnl.gov

Jackson B. Davidson
Oak Ridge National Laboratory
P. O. Box 2008
Oak Ridge, TN 37831-6010
615/574-5599

Ted Davis
New England Biolabs, Inc.
32 Tozer Road
Beverly, ME 01915
508/927-5054
Fax: 508/921-1350

Larry Deaven
Los Alamos National Laboratory
LS-4, MS-M888
Los Alamos, NM 87545
505/667-3114

S. B. Dev
BioTechnologies/Experimental Research
3742 Jewell Street
San Diego, CA 92109
619/270-0861; Fax: 619/483-3817

Jeanne Dietz-Band
Los Alamos National Laboratory
LS-3 MS M886, P. O. Box 1663
Los Alamos, NM 87545
505/667-2690
Fax: 505/665-3024

Norman Doggett
Los Alamos National Laboratory
MS M886
Los Alamos, NM 87545
505/665-4007
doggett@flovax.lanl.gov

Richard J. Douthart
Pacific Northwest Laboratory
P. O. Box 999, MS K4-13
Richland, WA 99352
509/375-2653
Fax: 509/375-6821
dick@gnome.pnl.gov

Radoje Drmanac
Argonne National Laboratory
Biological & Medical Research Division
9700 South Cass Avenue
Argonne, IL 60439-4833

Michele Durand
Scientific Attache
Embassy of France
4101 Reservoir Road
Washington, DC 20007
202/944-6235

Gary A. Epling
University of Connecticut
Department of Chemistry, U-60
Storrs, CT 06269
203/486-3215
Fax: 203/486-2981

Glen Evans
1445 Pesaso Street
Encinitas, CA 92024
619/453-4100
Fax: 619/558-9513
gevans@salk

Eric Fairfield
Los Alamos National Laboratory
Group T-10, MS K710
Los Alamos, NM 87545
505/665-0479
frf%life@lanl.gov

Brian Faldasz
State University of New York
Health Science Center at Syracuse
Department of Medicine
750 East Adams Street
Syracuse, NY 13210
315/464-5446

Steve Ferris
Bio-Rad Laboratories
15111 San Pablo Avenue
Richmond, CA 94806

Anne Fertitta
Human Genome Center
Lawrence Livermore National Laboratory
P. O. Box 5507, L-452
Livermore, CA 94551
415/423-3633
Fax: 415/423-3608

James W. Fickett
Los Alamos National Laboratory
Group T-10, MS K710
Los Alamos, NM 87545
505/665-5340
jwf%life@lanl.gov

Chris Fields
New Mexico State University
Box 30001/3CRL
Las Cruces, NM 88003-OO01
505/646-2848
Fax: 505/646-6218
cfields@nmsu.edu

Robert S. Foote
Oak Ridge National Laboratory
Biology Division
P. O. Box 2009
Oak Ridge, TN 37831-8077

Gerald Friedman
Los Alamos National Laboratory
MS MT08
Los Alamos, NM 87545
505/667-1447
Fax: 665-4408
gf@beta.lanl.gov

David Galas
Director, OHER-DOE
Washington, DC 20545
301/353-3251
Fax: 301/353-5051

Emilio Garcia
Human Genome Center
Lawrence Livermore National Laboratory
P. O. Box 5507, L-452
Livermore, CA 94551
415/422-8002
Fax: 415/423-3608

Raymond F. Gesteland
University of Utah
HHMI/Human Genetics Department
743 Wintrobe Building
Salt Lake City, UT 84112
801/581-5190
rayg@utahmed

Jeff Gingrich
Human Genome Center
Lawrence Berkeley Laboratory
MS 74-3110
One Cyclotron Road
Berkeley, CA 97420
415/486-6549
Fax: 415/486-6816

Alexander N. Glazer
University of California
Stanley/Donner ASU
229 Stanley Hall
Berkeley, CA 94720
Fax: 415/643-9290

Walter Goad
Los Alamos National Laboratory
MS K710
Los Alamos, NM 87545
505/455-2464
wbg@lanl.gov

Gerald Goldstein
Physical & Technical Research Division
OHER, DOE, ER 74-GTN
Washington, DC 20545

Deborah L. Grady
Los Alamos National Laboratory
Genetics Group, MS M886
Los Alamos, NM 87545
505/667-2695

Donald W. Graumann
General Atomics
P. O. Box 85608
San Diego, CA 92186-9784
619/455-4537
Fax: 619/455-4215

Joe Gray
Lawrence Livermore National Laboratory
Biomedical Sciences Division
P.O. Box 5507
Livermore, CA 94550
415/422-5610
Fax: 415/422-2282

Sharon L. Gray
Los Alamos National Laboratory
P. O. Box 1663, MS A114
Los Alamos, NM 87545
505/667-1600
Fax: 505/665-3858

Alan Greener
Stratagene
1109 North Torrey Pines Road
La Jolla, CA 92037

Albert Haces
Life Technologies, Inc.
Research Products Division
P.O. Box 6009
Gaithersburg, MD 20877

Peter Hahn
State University of New York
Health Science Center at Syracuse
Syracuse, NY 13210
315/464-5956

James F. Hainfeld
Brookhaven National Laboratory
Upton, NY 11973
516/282-3372
Fax: 516/282-3407

Mark L. Hammond
Los Alamos National Laboratory
MS M-888, LS-4
Los Alamos, NM 87545
505/665-3743
Fax: 505/665-3024

Carol Harger
Los Alamos National Laboratory
Group T-10, MS K710
Los Alamos, NM 87545
505/665-2881
cah%life@lanl.gov

Reece Hart
Salk Institute
Molecular Genetics Laboratory
10010 North Torrey Pines Road
La Jolla, CA 92037
617/453-4100 x451
Fax: 617/558-9513
hart@salk-scz.sdsc.edu

Fred C. Hartman
Oak Ridge National Laboratory
Biology Division
Oak Ridge, TN 37831-8080
FTS 624-0212
Fax: 624-9297

John R. Hartman
Computational Biosciences, Inc.
P. O. Box 2090
Ann Arbor, MI 48106
313/995-3560

Philip Hempfner
Los Alamos National Laboratory
Group T-10, MS K710
Los Alamos, NM 87545
505/665-5340

Gary Hermanson
Salk Institute
10010 North Torrey Pines Road
La Jolla, CA 92037
619/453-4100 x536
Fax: 619/558-9513

C. Edgar Hildebrand
Los Alamos National Laboratory
Genetics Group, MS M885
Los Alamos, NM 87545
505/667-2746
Fax: 505/665-3024
ceh@telomere.lanl.gov

Jeff Himawan
Harvard Medical School
Department of Biological Chemistry &
Molecular Biology
240 Longwood Avenue
Boston, MA 02115

Diane Hinton
Howard Hughes Medical Institute
6701 Rockledge Drive
Bethesda, MD 20817
301/571-0282

Lee Hood
California Institute of Technology
Biology Department
Pasadena, CA 91135
818/397-2762
hood@caltech.edu

Eliezer Huberman
Argonne National Laboratory
Biological & Medical Research Division
9700 South Cass Avenue
Argonne, IL 60439-4833
708/972-3819
Fax: 708/972-3387

Tim Hunkapiller
California Institute of Technology
139-74
Pasadena, CA 91125
Fax: 818/793-4627
tim@hood.caltech.edu

Andrew P. Hunt
Los Alamos National Laboratory
LS-4, MS M888
Los Alamos, NM 87545

Marge Hutchinson
Human Genome Center
Lawrence Berkeley Laboratory
50 B/3238
One Cyclotron Road
Berkeley, CA 94720
415/486-4727
mshutchinson@lbl.gov

K. Bruce Jacobson
Oak Ridge National Laboratory
Biology Division
P. O. Box 2009
Oak Ridge, TN 37831-8077

Joe Jaklevic
Human Genome Center
Lawrence Berkeley Laboratory
MS 70A-3363
One Cyclotron Road
Berkeley, CA 94720
415/486-5647
Fax: 415/486-5401

J.H. Jett
Los Alamos National Laboratory
MS M888
Los Alamos, NM 87545
505/667-3843
Fax: 505/665-3024
jett@flovax.lanl.gov

Bertrand Jordan
Lawrence Berkeley Laboratory
Stanley Hall
Berkeley, CA 94720
415/642-0348
Fax: 415/643-9290

Fa-Ten Kao
Eleanor Roosevelt Institute
1899 Gaylord Street
Denver, C0 80206
303/333-4515
Fax: 303/333-8423

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

Joe Katz
Lawrence Berkeley Laboratory
MS70A-4475A
One Cyclotron Road
Berkeley, CA 94720
415/486-5636

Stuart Kauffman
Santa Fe Institute
1120 Canyon Road
Santa Fe, NM
505-984-8800

Richard Keller
Los Alamos National Laboratory
MS G738
Los Alamos, NM 87545
505/667-3018

William Kolbe
Human Genome Center
Lawrence Berkeley Laboratory
MS 70A-2205
One Cyclotron Road
Berkeley, CA 94720
415/486-7199

Rajendra Krishnan
Washington University
School of Medicine
Box 8230, 660 South Euclid Avenue
St. Louis, MO 63110-1093
314/362-2771
Fax: 314/362-1232
krishnan@borcim.wustl.edu.

Jane Lamerdin
Lawrence Livermore National Laboratory
P. O. Box 5507
Livermore, CA 94550

Michael Lane
State University of New York
Health Science Center at Syracuse
Department of Medicine
750 East Adams Street
Syracuse, NY 13210
315/464-5446

Roger Laskin
Life Technologies, Inc.
Research Products Division
Molecular Biology R&D
P. O. Box 6009
Gaithersburg, MD 20877
301/670-8350
Fax: 301/921-2116

Eugene Lawler
University of California, Berkeley
1121 Oxford Street
Berkeley, CA 94707
415/642-4019
Fax: 415/642-4775
lawler@arpa.berkeley.edu

Charles B. Lawrence
Baylor College of Medicine
Department of Cell Biology
Molecular Biology Information Research
One Baylor Plaza
Houston, TX 77030
713/798-6226
Fax: 713/790-1275
chas@mbir.bcm.tmc.edu

Leonard Lerman
Massachusetts Institute of Technology
Biology Department, 56-743
77 Massachusetts Avenue
Cambridge, MA 02139
617/253-6658

Frederick C. Leung
Battelle Pacific Northwest Laboratory
P. O. Box 999, MS K4-13
Richland, WA 99352
509/375-2169
Fax: 509/375-6821

Suzanna Lewis
Human Genome Center
Lawrence Berkeley Laboratory
MS 50B-1123
One Cyclotron Road
Berkeley, CA 94720
415/486-7370
selewis@lbl.gov

Hwa A. Lim
Florida State University
467 Supercomputer Computations
Research Institute, B-186
Tallahassee, FL 32306-4052
904-644-7046
Fax: 904-644-0098
hlim@scri.fsu.edu.

Jonathan Longmire
Los Alamos National Laboratory
MS M886, LS-3, Genetics Group
Los Alamos, NM 87545
505/667-8208

Vladimir Makarov
University of Michigan
Biophysics Research Division
2200 Bonisteel Boulevard
Ann Arbor, MI 48109-2099
313/264-5258
Fax: 313/264-3233

Wlodek Mandecki
Abbott Laboratories
Corp. Molecular Biology D93D
Abbott Park, IL 60064
708/937-2236
Fax: 708/688-6046

Betty K. Mansfield
Oak Ridge National Laboratory
Human Genome Management Info. Sys.
P. O. Box 2008
Oak Ridge, TN 37831-6050
615/576-6669
Fax: 615/574-9888
bkq@ornl.gov

Jen-I Mao
Collaborative Research, Inc.
Human Genetics Research
Two Oak Park
Bedford, MA 01730
617/275-0004 x142
Fax: 617/891-5062

Babetta L. Marrone
Los Alamos National Laboratory
LS-4, M888
Los Alamos, NM 87544
505/667-3279
Fax: 505/665-3024
marrone@flovax.lanl.gov

Christopher Martin
Tropix, Inc.
47 Wiggins Avenue
Bedford, MA 01730
617/271-0045
Fax: 617/275-8581

John C. Martin
Los Alamos National Laboratory
P. O. Box 1663
Los Alamos, NM 87545

Richard Mathies
University of California
Department of Chemistry
Berkeley, CA 94720
415/642-4192

Mary Kay McCormick
Los Alamos National Laboratory
M886
Los Alamos, NM 87545
505/665-4438

Pat Medvick
Los Alamos National Laboratory
MS J580
Los Alamos, NM 87545
505/667-2676
pm@lanl.gov

Mortimer L. Mendelsohn
Lawrence Livermore National Laboratory
Biomedical Sciences Division
P. O. Box 5507, L-452
Livermore, CA 94551
415/422-5765
Fax: 415/422-2282

George Michaels
National Institutes of Health
Division of Comput. Research
Building 12A, Room 2051
Bethesda, MD 20892
301/402-1140
Fax: 301/402-0007
michaels@helix.nih.gov

Harvey Mohrenweiser
Human Genome Center
Lawrence Livermore National Laboratory
P. O. Box 5507, L-452
Livermore, CA 94551
415/423-0534
Fax: 415/422-2282

Fred A. Morse
Los Alamos National Laboratory
P. O. Box 1663, MS A114
Los Alamos, NM 87545
FTS 855-3858
Fax: 505/665-3858

Robert Mortimer
University of California
Department of Molecular & Cell Biology
One Cyclotron Road
353 Donner
Berkeley, CA 94720
415/643-8877
Fax: 415/642-8589

Robert K. Moyzis
Los Alamos National Laboratory
CHGS/MS M885
Los Alamos, NM 87545
505/667-3912

Richard Mural
Oak Ridge National Laboratory
Biology Division
P. O. Box 2009
Oak Ridge, TN 37831-8077
615/576-2938
Fax: 615/574-1274

David L. Nelson
Baylor College of Medicine
Institute for Molecular Genetics
One Baylor Plaza, T821
Houston, TX 77030
713/798-4787
Fax: 713/798-7383
nelson@condor.mbir.bcm.tmc.edu

David O. Nelson
Human Genome Center
Lawrence Livermore National Laboratory
P. O. Box 5507, L-452
Livermore, CA 94551
415/423-8898
Fax: 415/423-3608
daven@gauss.llnl.gov

Debi Nelson
Los Alamos National Laboratory
Group T-10, MS K710
Los Alamos, NM 87545
505/665-2594
debi%life@lanl.gov

Frank Olken
Human Genome Center
Lawrence Berkeley Laboratory
MS 50B-3220
One Cyclotron Road
Berkeley, CA 94720
415/486-5891

Anne Olsen
Human Genome Center
Lawrence Livermore National Laboratory
P. O. Box 5507, L-452
Livermore, CA 94551
415/423-4927
Fax: 415/423-3608
olsen@ecor1.llnl.gov

Elizabeth T. Owens
Oak Ridge National Laboratory
Human Genome & Toxicology Group
P. O. Box 2008
Oak Ridge, TN 37831-6050
615/574-0601
Fax: 615/574-9888
tug@ornl.gov

Ross Overbeek
Argonne National Laboratory
MCS 221/D236
9700 South Cass Avenue
Argonne, IL 60439
708/972-7856
Fax: 708/972-5986

Peter L. Pearson
1830 East Monument Street
Baltimore, MD
301/955-9705
Fax: 301/955-0054
pearson@welch.jhu.edu

Robert Pecherer
Los Alamos National Laboratory
Group T-10, MS K710
Los Alamos, NM 87545
505/665-1970
rmp%life@lanl.gov

Joanne E. Pelkey
Pacific Northwest Laboratory
P. O. Box 999, Mail Stop K1-86
Richland, WA 99352
509/375-6947
Fax: 509/375-3641
jo@gnome.pnl.gov

Roger B. Perkins
Los Alamos National Laboratory
P. O. Box 1663, MS A114
Los Alamos, NM 87545
FTS 855-3858
Fax: 505/665-3858

Theodore T. Puck
Eleanor Roosevelt Institute
for Cancer Research
1899 Gaylord Street
Denver, CO 80206

Dietmar Rabussay
Life Technologies, Inc.
Molecular Biology R&D
P. O. Box 9418
8717 Grovemont Circle
Gaithersburg, MD 20898
301/670-8332
Fax: 301/921-2116

J. M. Ramsey
Oak Ridge National Laboratory
P. O. Box 2008
Oak Ridge, TN 37831-6142
615/574-5662

Venigalla B. Rao
The Catholic University of America
Department of Biology
Washington, DC 20064
202/319-5271
201/319-5721

Robert Ratliff
Los Alamos National Laboratory
Los Alamos, NM 87545
505/667-2872

David E. Reichle
Oak Ridge National Laboratory
Environmental, Life & Social Sciences
P. O. Box 2008
Oak Ridge, TN 37831-6253
615/574-4333
Fax: 615/576-2912

Orly Reiner
Baylor College of Medicine
Institute for Molecular Genetics
One Baylor Plaza
Houston, TX 77030
713/798-4774
Fax: 713/798-7383

Arthur D. Riggs
Beckman Research Institute
City of Hope
Biology Department
1450 East Duarte Road
Duarte, CA 91010-0269
818/357-9711
Fax: 714/593-9339

Robert Robbins
National Science Foundation
Instrumentation & Resources Division
1800 G Street NW
Washington, DC 20550
202/357/9880
Fax: 202/357-7745
rrobins@note.nsf.gov

Randy Roberts
Los Alamos National Laboratory
MS J580
Los Alamos, NM 87545
505/665-4485

Elise Rose
Perkin Elmer Cetus Corporation
1400 53 Street
Emeryville, CA 94608-2997
415/420-3379
Fax: 415/547-2273

Stanley D. Rose
Perkin Elmer Cetus Corporation
1400 53rd Street
Emeryville, CA 94608
415/420-4218
Fax: 415/547-2273

Miguel Salmeron
Human Genome Center
Lawrence Berkeley Laboratory
One Cyclotron Road, MS 66
Berkeley, CA 94720
415/486-6230

Arbansjit K. Sandhu
University of Medicine and Dentistry
Department of Microbiology &
Molecular Genetics
185 South Orange Avenue
Newark, NJ 07109
201/456-3409
Fax: 201/456-3644

Karen Schenk
Los Alamos National Laboratory
Group T-10, MS K710
Los Alamos, NM 87545
505/665-3804
khs%life@lanl.gov

Eric Schmitt
Massachusetts Institute of Technology
Biology Department, 56-743
77 Massachusetts Avenue
Cambridge, MA 02139
617/253-6658

Stanley M. Schwartz
Chi Systems, Inc.
Gwynedd Plaza II
Spring House, PA 19477
215/542-1400
Fax: 215/542-1412
stan@grasp.cis.upenn.edu

David Searls
732 Yale Avenue
Swarthmore, PA 19081
215/648-2146
dbs@prc.unisys.com

Lauren Sears
New England Biolabs, Inc.
32 Tozer Road
Beverly, MA 01915
508/927-5054
Fax: 508/921-1350

Jude Shavlik
University of Wisconsin
Computer Science
1210 West Dayton Street
Madison, WI 53706
608/262-7784
Fax: 608/262-9777
shavlik@cs.wisc.edu

James Shero
Baylor College of Medicine
Institute for Molecular Genetics
One Baylor Plaza
Houston, TX 77030
713/798-4774
Fax: 713/798-7383

Hiroaki Shizuya
California Institute of Technology
Division of Biology, 147-75
Pasadena, CA 91125

Wigbert J. Siekhaus
Lawrence Livermore National Laboratory
P. O. Box 808, L-357
Livermore, CA 94551
415/422-6884
Fax: 415/423-7040
siekhaus@cmsl.llnl.gov

Paul H. Silverman
Beckman Instruments, Inc.
Box 3100
2500 Harbor Boulevard
Fullerton, CA 92634-3100
714/773-7745
Fax: 714/773-7637

Daniel J. Simpson
Los Alamos National Laboratory
LS-4, MS-M888
Los Alamos, NM 87545
505/665-3859
Fax: 505/665-3024

Thomas Slezak
Human Genome Center
Lawrence Livermore National Laboratory
P. O. Box 5507, L-452
Livermore, CA 94550
415/422-5746
Fax: 415/423-3608
tom@yac.llnl.gov

Cassandra Smith
Human Genome Center
Lawrence Berkeley Laboratory
One Cyclotron Road
Berkeley, CA 94720
415/643-6376
Fax: 415/642-1188

Lloyd Smith
University of Wisconsin, Madison
Department of Chemistry
Madison, WI 53706
608/263-2594
Fax: 608/262-0381
smith@bert.wisc.edu

Carol A. Soderlund
New Mexico State University
Box 30001/3CRL
Las Cruces, NM 88001
505/646-6430
cari@nmsu.edu

Joseph Sorge
Chairman and CEO
Stratagene
1109 North Torrey Pines Road
La Jolla, CA 92037

Sylvia Spengler
Human Genome Center
Lawrence Berkeley Laboratory
One Cyclotron Road, MS1-213
Berkeley, CA 94720
415/486-5874
Fax: 415/486-5717
sylviaj@violet.berkeley.edu

J. N. Spuhler
Los Alamos National Laboratory
Genetics Group
Los Alamos, NM 87545
505/982-0696

Raymond Stallings
Los Alamos National Laboratory
LS-3, MS M886
P. O. Box 1663
Los Alamos, NM 87545
505/667-2690
Fax: 505/665-3024

Marvin Stodolsky
U.S. Department of Energy
OHER, ER-72 GTN
Washington, DC 20545-0001
301/353-4475
Fax: 301/353-5051
stodolsky@oerv01er.doe.gov

Linda D. Strausbaugh
The University of Connecticut
Department of Molecular & Cell Biology
Box U-125
75 North Eagleville Road
Storrs, CT 06269-3125
203/486-2693
molce12@uconnvm

F. William Studier
Brookhaven National Laboratory
Biology Department
Upton, NY 11973
516/282-3390
Fax: 516/282-3407

Betsy Sutherland
Brookhaven National Laboratory
Upton, NY 11973
516/282-3293
Fax: 516/282-3407

Grant Sutherland
Adelaide Children's Hospital
Dept. of Cytogenetics & Mol.
72 King William Road
North Adelaide S.A. 5006
618/267-7284
Fax: 618/267-7342

Stanley Tabor
Harvard Medical School
Dept. of Biological Chemistry & Mol.
240 Longwood Avenue
Boston, MA 02115
617/432-3128
Fax: 617/738-0516

Ed Theil
Lawrence Berkeley Laboratory
One Cyclotron Road
MS 46A-1120
Berkeley, CA 94720
415/486-7501

David Thomassen
Lovelace, Inhalation Toxicology
P. O. Box 5890
Albuquerque, NM 87185
505/844-3108
Fax: 505/844-5403

David A. Thurman
Pacific Northwest Laboratory
P. O. Box 999, MS K1-86
Richland, WA 99352
509/375-6950
Fax: 409/375-3641
dave@gnome.pnl.gov

David Torney
Los Alamos National Laboratory
T-10, MS K710
Los Alamos, NM 87545
dct@life.lanl.gov

Barbara Trask
Human Genome Center
Lawrence Livermore National Laboratory
P. O. Box 5507, L-452
Livermore, CA 94551
415/422-5706
Fax: 415/422-2282

James Trebes
Lawrence Livermore National Laboratory
L-473
Livermore, CA 94551

Katherine Tynan
Human Genome Center
Lawrence Livermore National Laboratory
P. O. Box 5507, L-452
Livermore, CA 94551
415/423-7831
Fax: 415/422-2282

Don Uber
Lawrence Berkeley Laboratory
MS74-349
One Cyclotron Road
Berkeley, CA 94720
415/486-6378
Fax: 415/486-6816
uber@petvax.lbl.gov

Edward Uberbacher
Oak Ridge National Laboratory
Biology Division
P. O. Box 2009
Oak Ridge, TN 37831
615/574-1210
Fax: 615/574-1274

Jean-Michel Vos
School of Medicine
Lineberger Comp. Cancer Research
Chapel Hill, NC 27599-7295
919/966-6887
Fax: 919/966-3015
vox@med.unc.edu

Mark Wagner
Lawrence Livermore National Laboratory
P. O. Box 808, L-156
Livermore, CA 94551
415/422-2866
mwagner@kooler.llnl.gov

Robert Wagner
Los Alamos National Laboratory
Los Alamos, NM 87545

Ronald A. Walters
Los Alamos National Laboratory
P. O. Box 1663, MS A114
Los Alamos, NM 87545
FTS 855-3858
Fax: 505/665-3858

Denan Wang
Human Genome Center
Lawrence Berkeley Laboratory
529 Stanley Hall
One Cyclotron Road
Berkeley, CA 97420
415/642-5841
Fax: 415/642-1188

Gan Wang
The University of Connecticut
Box U-131
354 Mansfield
Storrs, CT 06269-2131

Robert Weiss
University of Utah
HHMI-Human Genetics Department
743 Wintrobe Building
Salt Lake City, UT 84112
801/581-5190
weiss@utahmed

Sherman Weissman
Yale University School of Medicine
Department of Human Genetics
SHM, Room II-126
333 Cedar Street
New Haven, CT 06510

Burton Wendroff
Los Alamos National Laboratory
B-284
Los Alamos, NM 87545
505/667-6497
bbw@lanl.gov

Thomas Whaley
Los Alamos National Laboratory
Group LS-2, MS M880
Los Alamos, NM 87545
505/667-2765

Mark Wilder
Los Alamos National Laboratory
LS-4, M888
Los Alamos, NM 87545
505/667-2750
wilder@flovax.lanl.gov

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

Jan A. Witkowski
Cold Spring Harbor Laboratory
Banbury Center
Cold Spring Harbor, NY 11724
516/549-0507
Fax: 516/549-0672

Judy M. Wyrick
Oak Ridge National Laboratory
P. O. Box 2008
Oak Ridge, TN 37831-6050
615/574-7781
Fax: 615/574-9888
bkq@ornl.gov

Michael S. Yesley
Los Alamos National Laboratory
P.O. Box 1663, MS A187
Los Alamos, NM 87545
505/665-2523
yesley_michael_s@ofvax.lanl.gov

Edward S. Yeung
Iowa State University
Department of Chemistry
Ames Laboratory
Ames, IA 50011
515/294-8062
Fax: 515/294-0105

Kaoru Yoshida
Lawrence Berkeley Laboratory
529 Stanley Hall
One Cyclotron Road
Berkeley, CA 94720
415/642-5841
Fax: 415/642-1188

Phil Youderian
CIBR
11099 North Torrey Pines Road
La Jolla, CA 92037
619/535-5471
Fax: 619/535-5472

Jing-Wei Yu
Eleanor Roosevelt Institute
1899 Gaylord Street
Denver, CO 80206
303/333-4515
Fax: 303/333-8423

Yiwen Zhu
Human Genome Center
Lawrence Berkeley Laboratory
MS74-362
One Cyclotron Road
Berkeley, CA 94720
415/486-7278

Manfred Zorn
Human Genome Center
Lawrence Berkeley Laboratory
50B-3216
One Cyclotron Road
Berkeley, CA 94720
415/486-5041