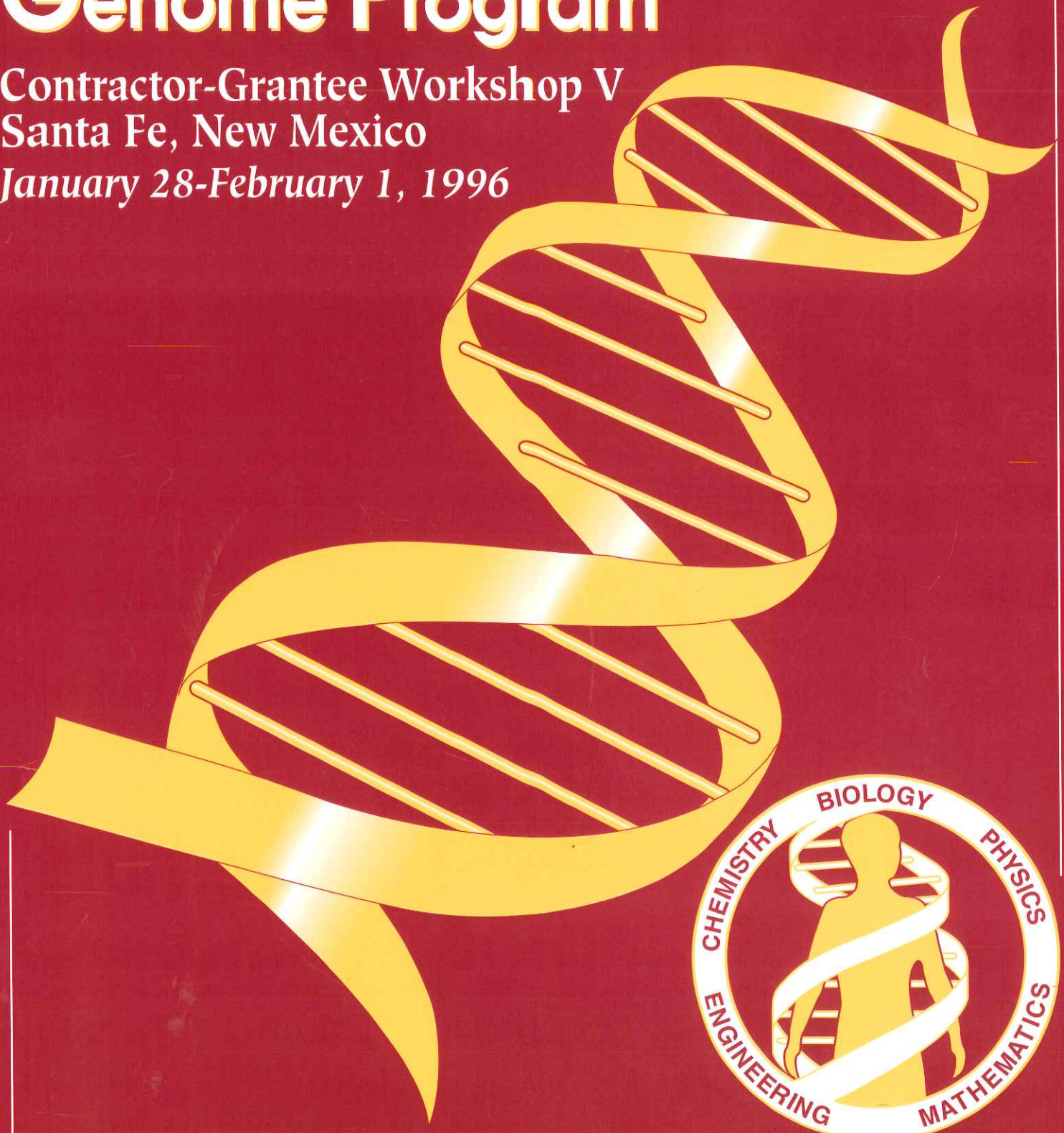
CONF-960143

DOE Human Genome Program

Contractor-Grantee Workshop V

Santa Fe, New Mexico

January 28-February 1, 1996



Please address queries on this publication to:

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

CONF-960143

DOE Human Genome Program Contractor-Grantee Workshop V January 28–February 1, 1996 Santa Fe, New Mexico

---

Date Published: January 1996

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 37830-6480

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

Contents

DOE Human Genome Program Contractor–Grantee Workshop V

Santa Fe, New Mexico
January 28–February 1, 1996

Plenary sessions are in the Eldorado Hotel; poster sessions and computer demonstrations are in the La Fonda Hotel. Each speaker and demonstration in the plenary sessions will have an abstract number and a poster associated with the oral presentation. Schedule correct as of December 15, 1995. Agenda is subject to change.

PLATFORM PRESENTATIONS
TENTATIVE SCHEDULE

Sunday, January 28, 1996

6:00–8:30p	Reception

Monday, January 29, 1996

8:00a	Opening Notes **A. Patrinos, M. Krebs**
Session I	**A. Carrano, Chair**
8:30a	C. Bult, TIGR
8:55	M. Simon, Caltech
9:20	G. Keen, NCGR
9:45	M.B. Soares/K. Elliston, Columbia U./Merck
10:15	Break
10:45	R. Mathies/A. Glazer, UC Berkeley
11:10	R. Eisenberg, U. of Michigan, Ann Arbor
11:35	L. Smith, U. of Wisconsin, Madison
12:00	L. Rowen, U. of Washington, Seattle
12:25–1:30p	Lunch
1:30-2:30	Poster Setup

Session II	Sequence Assembly Discussion **D. Kingsbury, Moderator**
2:30	P. Green, U. of Washington, Seattle
2:50	G. Sutton, TIGR
3:10	S. Pitluck, LBL
3:30	G. Myers, U. of Arzonia, Tucson
3:50	Discussion
4:45–7:00	Poster Session & Computer Demonstrations

Tuesday, January 30, 1996

Session III	**R. Moyzis, Chair**
8:00a	H.-U. Weier, LBL
8:25	J. Lamerdin, LLNL
8:50	E. Yeung, Ames Labs
9:15	C. Cantor, Boston U.
9:40	C. Martin, LBL
10:05	R. Weiss, U. of Utah, Salt Lake City
10:30	Break
11:00	M. Sosa, AAAS
11:25	G. Stormo, U. of Colorado, Boulder
11:50	B. Karger, Barnett Institute
12:15p	Lunch
Session IV	**L. Smith, Chair**
3:00	M. Graves, Baylor College
3:20	J. Ju, UC Berkeley
3:40	M. Shannon, ORNL
4:00	C. Boysen, U. of Washington, Seattle
4:25	R. Smith, Baylor College
5:15–7:30	Poster Session & Computer Demonstrations

Wednesday, January 31, 1996

Session V — **R. Gesteland, Chair**

8:00a	M. Quesada, BNL
8:25	K. Fasman, GDB
8:50	J. Dunn, BNL
9:15	B. Scott, Genome Radio Project
9:40	R. Smith, PNL
10:05	Break
10:30	P. de Jong, Roswell Park Cancer Center
10:55	H.-C. Chi, LANL
11:20	J. Korenberg, Cedars-Sinai, UCLA
11:45	E. Uberbacher, ORNL
12:10p	M.A. Brow, Third Wave Technologies
12:35	Lunch

Session VI — **M. Narla, Chair**

3:00	F. Zweig, Einstein Institute
3:25	G. Church, Harvard U.
3:50	L. Stubbs, ORNL
4:15	T. Hawkins, Whitehead Institute
5:00–7:30	Poster Session & Computer Demonstrations

Thursday, February 1, 1996

Session VII — **D. Kingsbury, Chair**

8:00a	J. McInerney, BSCS
8:25	D. Nickerson, U. of Washington, Seattle
8:50	M. Pirrung, Duke U.
9:15	E. Eichler, LLNL
9:40	D. Torney, LANL
10:05	Break
10:35	E. Garcia, LLNL
11:00	D. Ricke, LANL
11:25	N. Dovichi, U. of Alberta, Edmonton
12:00noon	Lunch

Introduction to Contractor-Grantee Workshop V

Welcome to the Fifth Contractor-Grantee Workshop sponsored by the Department of Energy (DOE) Human Genome Program (HGP). This meeting offers an important opportunity for investigators and program managers to review the program's progress and content, to assess the impact of new technologies and, not least, for scientists to forge new collaborations.

The 197 abstracts in this book describe the research of DOE-funded grantees and contractors from DOE's human and microbial genome programs (plus the work of a few invited guests). These abstracts represent the most up-to-date compilation of activities and accomplishments in these fast-moving programs.

Many projects will be discussed at plenary sessions to be held at the Eldorado Hotel, and all projects are represented by posters. Space has been set aside at the La Fonda Hotel for displaying posters and demonstrating new informatics resources.

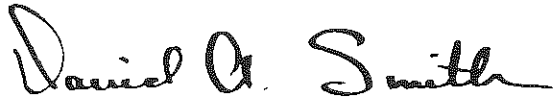
The U.S. Human Genome Project officially reached its fifth birthday on October 1, 1995. In the last 5 years, remarkable progress has been made toward the goals that were laid down in the original and revised 5-year plans that have guided the project. At this workshop we will hear about many of the significant advances that have been made since we last met in November 1994. We remain ahead of schedule and under budget, and there is reason to believe that the project will reach its goals before the target date of 2005. Nevertheless, major challenges remain.

Obtaining the complete human DNA sequence is a monumental task that we are just beginning to address seriously; it will continue to require a very large, focused, and dedicated effort. However, in the last several months there has been growing optimism that the sequencing goal is now coming within reach. This optimism stems from recent successes in sequencing the genomes of simpler organisms, coupled with new and oncoming advances in sequencing reagents and instrumentation and the development of improved clone resources and algorithms that reduce human decision making. With the growing consensus that cost-effective approaches to large-scale DNA sequencing are nearly within our grasp, it is clear that to reach our sequencing goals, we must begin to commit a significantly larger fraction of the HGP budget into the focused effort necessary for production of large amounts of sequence data.

These are exciting and challenging times for biological researchers. The wealth of information and capabilities now being generated by the various genome projects and other biological endeavors will lead over the next two decades to more insights into living systems than have been amassed in the past two millennia. Biology truly is undergoing a revolution.

Finally, as most of you know, I will retire on January 26, 1996, just before the workshop. It has been my privilege to be involved with the Human Genome Project since it first became a glimmer in the minds of a few. I am very proud of the many significant contributions that have been made by DOE-supported genome researchers and would like to thank all of you for your work in moving these programs forward.

I am anticipating a very interesting and productive meeting and offer special thanks to all who have contributed to its organization.



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

December 19, 1995

Sequencing

This page intentionally left blank.

DIRECTED GENOMIC DNA SEQUENCING ON HUMAN CHROMOSOME 5

Christopher H. Martin, Cheryl A. Davis, Cheryl L. Ericsson, Carol A. Mayeda, Herb Moise and Michael J. Palazzolo
Human Genome Center
Lawrence Berkeley Laboratory, Berkeley, CA 94720

Our group has developed a novel directed approach to genomic sequencing in which every sequencing template is mapped to a resolution of 30 base pairs prior to being sequenced. This high resolution mapping information yields two important advantages. First, genomic sequence can be determined with far fewer sequencing reactions than approaches which utilize random coverage to yield the bulk of the complete sequence. Second, the difficulty of the sequence assembly problem is greatly reduced, as this high resolution physical mapping information provides significant guidance to the process of reconstruction of the original genomic sequence.

Using our directed approach, we have completed over 1,000,000 base pairs of completely double stranded and edited human genomic sequence. Our initial sequencing target in the human genome consists of a large growth factor rich region located at 5q31-q35. All of this sequence is derived from sequencing subclones derived from P1 physical mapping clone sets that span this region. We have found the P1 physical mapping system to be an excellent substrate for our genomic sequencing efforts. This megabase of sequence is all of high quality (less than 1 error per 2,500 base pairs and all completely double stranded) and was completed during the past two years. Based on these achievements, we are now in the process of scaling up our efforts in human genomic DNA sequencing at LBNL.

In addition to sequence production efforts, a major focus of the center is on technology development, where our goal is the continuing improvement of the directed sequencing strategy. In the area of biology, we have recently replaced an agarose/PCR based step in our high resolution physical mapping phase with one based on sequencing of the ends of several ~3 kb subclones. In automation, a new integrated platform for PCR reaction setup and subsequent thermocycling is the entering of the production testing phase. In the area of informatics, we are beginning to use new tools being developed both at the LBNL Center and also in collaboration with Gene Meyers at the University of Arizona. These tools capture and reflect the high resolution physical mapping information made available by the directed sequencing strategy for sequence assembly and automation of many sequence finishing decisions. Overall, we are continuing to make progress towards our goal of using the directed strategy as a biological platform for the development of a highly automated system for the sequencing of genomic DNA.

Computational and Biological Analysis of 1.2 Mb of sequence at 5q31

Kelly A. Frazer, Yukihiko Ueda, Maria R. Garofalo, Jan-Fang Cheng, Nomi Harris, Frank Eeckman, and Edward M. Rubin
Human Genome Center, Lawrence Berkeley National Laboratory, 1 Cyclotron Road MS 74-154, Berkeley, CA 94720

The LBNL Human Genome Center production sequencing group is focusing on the interleukin growth factor gene cluster region of chromosome 5q31. The human biology group is working in collaboration with the informatics group to determine the location and function of unknown genes which exist within this region. To date the production sequencing group has generated approximately 390 kb of non contiguous sequence in this 1.2 Mb region of chromosome 5q31. To determine the location of putative genes within this sequence data we have analyzed 240 kb of sequence using components of the informatics group's annotation workbench.[1] We found 83 potential exons employing the genefinding program GRAIL. To learn about the potential function of these 83 predicted exons we conducted sequence comparison searches using the BLAST program. The database searches demonstrated that 11 of these exons correspond to three genes previously known to lie in the region, IL13, IL3, and CSF2. Eight exons correspond to 2 previously unidentified human genes which are highly homologous to the mouse KIF3 and the rat phosphatidylinositol 4-kinase genes. Three exons correspond to EST sequences representing the 3' ends of 2 anonymous cDNA clones. The remaining 61 potential exons were not significantly homologous to any sequence in GSDB. Our computational analysis of 20% of the 1.2 Mb region of chromosome 5q31 has found exons corresponding to 3 of the 6 genes previously known to lie in the region, 2 previously unidentified human genes that are homologs of genes which have been studied in other species, and two new anonymous human genes.

One of our approaches to determine the biological function of these newly identified genes utilize YAC transgenic mice. It has been previously demonstrated that human genes harbored on YACs in transgenic mice are temporally and spatially correctly regulated. We have created 4 lines of transgenic mice which contain overlapping YACs from the 5q31 region. This set of transgenic mice which together harbor 1.8 Mb of contiguous human sequence serve as a substrate to assess the expression pattern and potential function of putative genes discovered in the 5q31 region. We have analyzed the temporal and spatial expression pattern of the human homolog of the mouse KIF3 gene. KIF3 is a kinesin-like protein involved in microtubule intra-cellular transport and has been shown by Northern blot analysis to be specifically expressed in the brain. RNase protection analysis of the human homolog of KIF3 in tissue isolated from the YAC transgenic mice was performed. This analysis demonstrated that the human KIF3 gene is expressed and transcribed preferentially in brain tissue. These studies demonstrate the usefulness of expression studies in YAC transgenic mice to biologically annotate large stretches of genomic sequence information.

[1]Nomi Harris and Frank Eeckman, A Workbench for Sequence Annotation and Browsing, please see abstract.

Sequencing by Parallel Primer Walking of a 240 kb Human Chromosome 7q Telomere Region: Sample Sequencing (SASE) Analysis as a Framework for Complete Large Scale Genomic Sequencing*

Han-Chang Chi, Judy M. Buckingham, A. Christine Munk, Elizabeth Saunders, Rebecca Lobb, Jingmei Liu, Quincey Simmons, Michael R. Altherr, Darrell Ricke, Jung-Rung Wu, and Robert K. Moyzis. Center for Human Genome Studies and Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545.

Most human telomeric genes are hundreds of kilobases from the chromosome termini, which consist of tandem arrays of the simple repeat sequence (TTAGGG)n[1]. Scrambled arrays of more complex repetitive DNA are located internal to the simple sequence repeat[2]. The telomeric end of chromosome 7q appears, however, to lack significant blocks of subtelomeric repetitive DNA[3]. In order to determine what human genes, if any, are near the 7q terminus, as well as whether or not those genes are affected by the telomeric DNA shortening demonstrated to occur in somatic cells, complete DNA sequencing of a 240 kb YAC, representing the end of this chromosome was performed.

Cosmid contigs representing the entire 240 kb terminal DNA from a human 7q telomeric yeast artificial chromosome clone (HTY-146) were constructed. From seven overlapping cosmids and a 11 kb DNA fragment amplified by polymerase chain reaction, a Sample Sequencing (SASE) approach that rapidly generates aligned sequences was initated (see Ricke et al., this meeting). "Nucleation points" from the cosmids or DNAs for SASE analysis were obtained by 1) end sequencing individual cosmids or DNAs, 2) subcloning of single restriction fragments and end sequencing, 3) sequencing Sau3AI total digest subclones, or 4) end sequencing partial digest random 3 kb subcloned fragments.

By assemblying and aligning these nucleation points, 70% sequence coverage is achieved with 98% clone coverage from a one-pass, 96-sample sequencing of a cosmid. From these initial sequences of the SASE clones, oligonucleotides were synthesized and further sequence data obtained by sequencing and primer walking directly off the original cosmid DNA. Both DNA sequence analysis and exon trapping experiments have identified some potentially interesting features dispersed along this terminal region of chromosome.

*This work was funded by the U. S. Department of Energy under contract W-7405-ENG-36.

References:

1. Moyzis, R. K. et al., *Proc. Natl. Acad. Sci. USA* **85**, 6622-6626 (1988).
2. Riethman, H. C. et al., *Proc. Natl. Acad. Sci. USA* **86**, 6240-6244 (1989).
3. Riethman, H. C. et al., *Genomics* **17**, 25-32 (1993).

SEQUENCING HUMAN CHROMOSOME 16: SAMPLE SEQUENCING (SASE) ANALYSIS AS A FRAMEWORK FOR IDENTIFYING GENES AND COPLETE LARGE-SCALE GENOMIC SEQUENCING*

Darrell O. Ricke, Judith M. Buckingham, A. Christine Munk, Rebecca Lobb, Elizabeth H. Saunders, Jingmei Liu, Norman A. Doggett, Michael R. Altherr, Larry L. Deaven, and Robert K. Moyzis. Life Sciences Division and Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545.

The human chromosome 16 physical map (Doggett *et al.*, Nature 377:Suppl:335-365, 1995; Doggett *et al.*, this meeting) provides the ideal framework for sequencing a human chromosome. We are using a SAmple SEquencing (SASE) approach to rapidly generate aligned sequences along the chromosome 16 physical map. SASE analysis is a method for rapidly "scanning" large genomic regions with minimal cost, identifying, and localizing most genes. Briefly, individual cosmids are partially digested with Sau3A and 3 kb fragments are recloned into double-strand sequencing vectors. By sequencing both ends of a 1X sampling of these recloned fragments along with end sequences of the cosmid, 70% sequence coverage is achieved with 98% clone coverage. The majority of this clone coverage is ordered by the relationship between the subclone end sequences. These ordered sequences are ideal substrates for directed sequencing strategies (see Chi *et al.*, this meeting). SASE analysis has been initiated on the 40 Mb short arm of chromosome 16. We propose to make chromosome 16 SASE sequences, along with feature annotation, publicly available through GSDB. Such data are sufficient to allow PCR amplification of the sequenced region from GSDB submissions alone, eliminating the need for extensive clone archiving and distributing. Therefore, SASE analysis provides the opportunity for numerous laboratories to complete the distributive genomic sequencing of chromosome 16. To identify and annotate regions of biological interest, we have developed the SCAN (Sequence Comparison ANalysis) algorithm to extract and identify significant homologies from database search results of SASE data. Initial SCAN results on the first 0.6 Mb of SASE data analyzed have identified multiple candidate genes and exons.

*This work is funded by USDOE under contract W-7405-ENG-36.

Cycle Sequencing with 18-mers Produced from a Hexamer Library by Ligation of Hexamers on Hexamer Templates

John J. Dunn, Laura-Li Butler-Loffredo and F. William Studier
Biology Department, Brookhaven National Laboratory, Upton, New York 11973

Primer walking using oligonucleotides selected from a library is an attractive strategy for large-scale DNA sequencing. Strings of three adjacent hexamers can prime DNA sequencing reactions specifically and efficiently when the template is saturated with a single-stranded DNA-binding protein (1), and a library of all 4,096 hexamers is manageable. We would like to be able to sequence directly on 35-kbp fesmid templates, but the signal from a single round of synthesis is relatively weak and triple-hexamer priming has not yet been adapted for cycle sequencing. We reasoned that a hexamer library might be used for cycle sequencing if combinations of hexamers could be selectively ligated by using other hexamers as the template for alignment. In this way, the longer primers needed for cycle sequencing could be generated easily and economically without the need for complex machines for *de novo* synthesis.

We found that ordered ligation of 3 hexamers to form an 18-mer occurs readily on a template of the 3 complementary hexamers (offset by three base pairs) that can base pair unambiguously to form a double-stranded complex of indefinite length (2). Each hexamer forms three complementary base pairs with two other hexamers, generating complementary chains of contiguous hexamers with strand breaks staggered by three bases. Two adjacent hexamers in the chain to be ligated contain 5' phosphate groups and the others are unphosphorylated. Both T4 and T7 DNA ligase can ligate the phosphorylated hexamers to their neighbors in such a complex at hexamer concentrations in the 50-100 μM range, producing an 18-mer and leaving three unphosphorylated hexamers. The products of these ligation reactions can be used directly for fluorescent cycle sequencing of 35-kbp templates.

Unambiguous ligation requires that alternative complexes with perfect base pairing not be possible with the combination of hexamers used. Since the combination of hexamers is dictated by the sequence of the desired ligation product, some oligonucleotides cannot be produced unambiguously by this method. However, 82.5% of all possible 18-mers could potentially be generated starting with a library of all 4096 hexamers, more than adequate for high throughput DNA sequencing by primer walking.

Supported by the Office of Health and Environmental Research of the U. S. Department of Energy.

(1) Kieleczawa, J., Dunn, J. J., and Studier, F. W. DNA sequencing by primer walking with strings of contiguous hexamers. *Science*, **258**, 1787-1791 (1992).

(2) Dunn, J. J., Butler-Loffredo, L. and Studier, F. W. Ligation of hexamers on hexamer templates to produce primers for cycle sequencing or the polymerase chain reaction. *Anal. Biochem.* **228**, 91-100 (1995)

Fesmid Vector for DNA Mapping and Sequencing

John J. Dunn, Matthew Randesi and F. William Studier
Biology Department, Brookhaven National Laboratory, Upton, New York 11973

We have developed a vector, referred to as a fesmid, for making libraries of approximately 35-kbp DNAs for mapping and sequencing. The high efficiency λ packaging system is used to generate libraries of clones. These clones are propagated at very low copy number under control of the replication and partitioning functions of the F factor, which helps to stabilize potentially toxic clones. A P1 lytic replicon under control of the *lac* repressor allows amplification simply by adding IPTG. The cloned DNA fragment is flanked by packaging signals for bacteriophage T7, and infection with an appropriate T7 mutant packages the cloned sequence into T7 phage particles, leaving most of the vector sequence behind. The size of the vector portion is such that genomic fragments packageable in λ (normal capacity 48.5 kbp) should also be packaged in T7 (normal capacity 40 kbp).

We have made fesmid libraries of several bacterial DNAs, including *Borrelia burgdorferi* (the cause of Lyme disease), *Bartonella henselae* (the cause of cat scratch fever), *E. coli*, *B. subtilis*, *H. influenzae*, and *S. pneumoniae*, some of which have been reported to be difficult to clone in cosmid vectors. Human DNA is also readily cloned in these vectors. Brief amplification followed by infection with a gene *3* and *17.5* double mutant of T7, which is defective in replicating its own DNA, produces lysates in which essentially all of the phage particles contain the cloned DNA fragment. Simple techniques yield high-quality DNA from these phage particles. Primers for direct sequencing from the ends of fesmid clones have been made.

Primer walking from the ends of fesmid clones could be an efficient way to sequence bacterial genomes, YACs, or other large DNAs without the need for prior mapping of clones. The ends of fesmids from a random library provide multiple sites to initiate primer walking. Merging of the elongating sequences from different clones will simultaneously generate the sequence of the original DNA and determine the order of the clones. The packaged fesmid DNAs are a convenient size for multiple restriction analyses to confirm the accuracy of the nucleotide sequence.

Supported by the Office of Health and Environmental Research of the U. S. Department of Energy.

Complete Sequence of a 34.8-kbp Fesmid Clone of *Borrelia burgdorferi* Determined in Part by Primer Walking with Hexamer Strings

John J. Dunn, Laura-Li Butler-Loffredo, Jan Kieleczawa and F. William Studier
Biology Department, Brookhaven National Laboratory, Upton, New York 11973

Borrelia burgdorferi is a spirochete that causes Lyme disease. It has a linear chromosome of about 935 kbp that contains about 70% AT base pairs. Although other workers have reported difficulty in cloning and maintaining long pieces of *Borrelia* DNA in *E. coli*, we readily obtained a library in our fesmid vector, presumably because the cloned DNA is propagated at very low copy number under control of the replication and partitioning functions of the F factor. Using a *Sau*3A partial digest of DNA from a very early passage strain, we picked 250 clones with inserts averaging about 35 kbp, an estimated 9-10 fold coverage of the *Borrelia* chromosome. We are beginning to sequence this chromosome as a test system in which to develop methods for sequencing DNA by primer walking using a hexamer library.

We have completed the sequence of one fesmid insert by generating both random and specific subclones in pGEM-based plasmid vectors, sequencing both ends of each subclone using standard vector primers, and filling the gaps and merging the entire sequence by primer walking with hexamer strings or by cycle sequencing with 18-mers formed by ligation of hexamers on hexamer templates. Many of the subclones were obtained by partial digestion with *Eco*RI under relaxed "star" conditions or with *Tsp*509 I, since both enzymes leave AATT overhangs that can be directly ligated into the *Eco*RI site of the pGEM vectors. *Bgl*II, *Hin*dII and *Pst*I sites were also used in generating the subclones. A gap not covered by the subclones was filled by primer walking directly on the fesmid DNA.

Both strands of this fesmid insert were completely sequenced, a total of 34,820 base pairs. Sequences determined by priming with hexamer strings or ligated hexamers cover more than 80% of the insert. This part of *Borrelia* DNA contains a high density of coding sequence (97%), with 36 complete open reading frames plus incomplete coding sequences at each end, all transcribed in the same direction. Similarities with genes from other bacteria indicate that operons for septum formation and biosynthesis of the basal body and flagellum are contained in this region of the *Borrelia* genome.

Supported by the Office of Health and Environmental Research of the U. S. Department of Energy and by a grant from the National Institute of Allergy and Infectious Diseases.

Automated Multiplex Microbial Genome Sequencing

Robert B. Weiss , Mark Stump, J oshua Cherry, Brett Duval , Robert Black , Sandy Kazuko, Jane Macfarlane, Diane Dunn, and Peter Cartwright.
Department of Human Genetics, University of Utah, Salt Lake City, UT 84112.

We have initiated large-scale genomic DNA sequencing of microorganisms of industrial and biological interest. Our first project is to complete the genome sequence of *Pyrococcus furiosus* (DSM 3638), a hyperthermophilic member of the Archaea. This organism, isolated from hot marine sediments, grows vigorously at temperatures near 100°C. Novel methods and instrumentation, developed at the Utah Center for Human Genome Research, are being used [1] to sequence this ~2.0 Mb genome. The contiguous sequences of individual plasmid inserts are determined by a multiplexed transposon-based directed strategy. The University of Utah Probe Chambers, automated devices for detecting enzyme-linked fluorescence from DNA hybrids on nylon membranes, are being used in both the mapping and sequencing phase of the project.

P. furiosus plasmid and cosmid libraries have been built in a family of 21 multiplex vectors. These vectors provide multiplex tags for both the end sequencing and transposon mapping phases of the process. Common growth and DNA prep formats feed both the mapping and sequencing process. Minimal spanning sets of clones are predicted from the mapping phase and fed to the sequencing process, where cycle sequencing is performed on multiplexed double-stranded templates. Multiplex sequence ladders are transferred to nylon membranes using a direct transfer apparatus, and then placed in the Probe Chambers for multiple cycles of data collection.

Currently, over 1.5 Mb of genomic DNA clones spread across 153 plasmid and cosmid inserts have entered, and are exiting, the mapping phase. The first bolus of mapped clones that exited the mapping phase comprised 2,518 mapped priming sites spanning 0.7 Mb of genomic clones. These sets of mapped priming sites are being rapidly sequenced using two Probe Chambers, each emitted 108 sequence reads every 10.5 hr.. This initial random coverage of the genome will be superceded in the closure phase by sequence matching between newly completed insert sequences and a database of end sequences from the plasmid and cosmid libraries. This directed selection of new inserts to be sequenced, will lead to completion of the genome if all the initial gaps are sequence gaps. The *Pyrococcus* project will enter the closure phase in the early 1996. This project is an alpha test of methods, instrumentation and software under development at the Utah Center for Human Genome Research.

This work is funded by DOE grant DE-FG03-94ER-61950 (R.B. Weiss, P.I.)

[1] Cherry, J.L., Young, H., Di Sera, L.J., Ferguson, F.M., Kimball, A.W., Dunn, D.M., Gesteland, R.F., and Weiss, R.B. (1994). Enzyme-linked fluorescent detection for automated multiplex DNA sequencing. *Genomics* **20**, 68-74

THE GENOME SEQUENCE OF *METHANOBACTERIUM THERMOAUTOTROPHICUM**

Douglas R. Smith, Lynn Doucette-Stamm, Hong Mei Lee, Joanne Dubois, Craig Deloughery, Tyler Aldredge, Romina Bashirzadeh, Derron Blakely, Wendy Caubet, Maria Chung, Katie Gilbert, Chenghua Ma, Pamela Parenteau, Rupal Patel, Dayong Qui, Skip Shimer, Xing Wang, Jamey Wierzbowski, Jork Nolling[†] and John Reeve[†], Genome Therapeutics Corp., 100 Beaver St. Waltham, MA 02154 [†]The Ohio State University, 484 W. 12th Ave., Columbus, OH, 43210

The goal of this project is to sequence the genomes of microbes which may be useful for energy production and bioremediation of toxic wastes. In related projects we are sequencing the genomes of bacterial pathogens and regions of human chromosome 10. The sequencing is being done by computer-assisted[‡] multiplex sequencing techniques which are under active development through an NIH-funded Genome Science and Technology Center.

Our initial focus has been the genome of the archaeon *Methanobacterium thermoautotrophicum* (1.7 Mb), which is ubiquitous in anaerobic environments and is potentially useful for the production of methane from biowastes. The organism can be readily grown and manipulated under laboratory conditions. The sequencing was done by a whole-genome shotgun approach with 2 kb plasmid subclones. Pools of templates were sequenced by chemical and enzymatic cycled-sequencing methods and run on direct transfer electrophoresis gels. A set of 67 nylon membranes containing reactions from 30,720 templates (1536 pools of 20 clones) were sequentially probed to generate films that were then scanned into a computer system. Approximately 13 Mb of raw data were generated (7.5 genome equivalents). The data was assembled into contigs using the program PHRAP and primers were automatically selected from the ends of the contigs using the program AUTOPRIMER. Dye-terminator finishing reactions were then performed and the samples run on ABI 377 machines. The data from the finishing reactions was then re-assembled together with the original data into approximately half the original number of contigs. This process was then reiterated to further reduce the number of contigs.

A preliminary analysis for database homologies was performed using the program BLAST on individual sequence reads. In-depth analysis will commence once sufficiently large contigs have been generated and have been proofread to reduce the occurrence of indels. The sequences will be made available with full annotation for gene locations as soon as possible after completion. The group at Ohio State University has provided starting DNA and cosmids, and is assisting in the analysis of the data.

*Supported under the Microbial Genome Program by Grant No. DE-FC02-95ER61967 from the Office of Health and Environmental Research of the US. Department of Energy

[‡]automated image analysis and sequence reading for multiplex sequencing is currently done using REPLICA™, a program developed by Mintz and Church at HHMI at Harvard Medical School

Use of Partial *Cvi*JI Digestion as an Alternative Approach to Generate Cosmid Sublibraries for Large-Scale DNA Sequencing Projects

Jeffrey C. Gingrich, Denise M. Boehrer, and Subha B. Basu*
Human Genome Center, Biology and Biotechnology Research Program, L-452,
Lawrence Livermore National Laboratory, 7000 East Ave., Livermore, CA 94551
*Present address: Genetic Systems Division, Life Science Group, Bio-Rad Laboratories,
2000 Alfred Nobel Drive, Hercules, CA, 94547

The generation of random fragments for shotgun sequencing projects is typically accomplished by pressure shearing or sonication. Digestion of the template DNA with restriction enzymes offers several advantages to these methods. Since little DNA is needed for optimization of the digestion and cloning, large scale preparations of cosmid DNA by cesium banding or other time consuming approaches are not necessary. Furthermore, since end-repair of the DNA is not necessary, a significantly higher cloning efficiency is seen following restriction enzyme digestion compared to sheared or sonicated DNA.

The restriction enzyme *Cvi*J1 recognizes between a two and three base DNA sequence. Complete digestion with *Cvi*JI thus results in DNA fragments averaging from 16 to 64 nucleotides in length. Partial digestion with *Cvi*JI can therefore fragment DNA in a "quasi" random fashion similar to shearing or sonication. In order to explore the use of *Cvi*JI for a large scale shotgun sequencing project, we have generated an M13 sublibrary using *Cvi*JI partial digestion on miniprep DNA from a previously sequenced human cosmid clone. This cosmid contains 1,031 "normal" (PuGCPy) *Cvi*JI sites with an average fragment size of 25 nucleotides. The DNA sequences for the first ~500 bp from 103 M13 subclones averaging 1.1 kb in size were determined. The DNA sequences from 3 of the subclones did not match sequences previously determined for the cosmid, indicating that miniprep DNA can used for libraries with little contamination with genomic *E. coli* or other DNA. DNA sequences from the other 100 subclones matched previously determined sequences of the cosmid. The cosmid coverage of each M13 subclone was estimated using the sequenced cloning site as the starting point and the estimated size of each insert. The 100 subclones were estimated to cover 90% of the cosmid. Since these subclones total 2.5 X in cosmid coverage, the cosmid coverage would have been 92% if these subclones were generated entirely at random. The use of *Cvi*J1 to generate random DNA fragments thus offers a simple alternative with many advantages compared to other commonly used fragmentation methods for the generation of random sublibraries in large scale shotgun DNA sequencing projects.

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

GENOMIC SEQUENCING OF HUMAN CHROMOSOME 19 AND COMPARATIVE ANALYSIS OF HUMAN AND RODENT DNA REPAIR GENE REGIONS.

Jane Lamerdin, Mishelle Montgomery, Stephanie Stilwagen, Melissa Ramirez, Aaron Adamson, Subha Basu, Ami Kyle, Paula McCready, Jeff Gingrich, Anne Olsen, Larry Thompson, Emilio Garcia, and Anthony Carrano. Human Genome Center, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA, 94550

Approximately 83% of human chromosome 19 is spanned by cosmids for which an EcoR1 map has been derived. We are scaling our sequencing facility to take advantage of these ordered clones to provide high-throughput, high-accuracy sequence for all of chromosome 19. In addition, we are utilizing our rapid clone selection/mapping capabilities to sequence other genomic regions associated with DNA repair and disease susceptibility. We have completed ~ 650 kbp of genomic sequence to high accuracy and have another ~500 kbp in various stages of completion (from assembly through annotation). Our primary effort has been targeted to cosmids containing the human DNA repair genes *HHR23A*, *XRCC1* and *ERCC2* on chromosome 19, *ERCC4* on chromosome 16, *XRCC3* on chromosome 14, and *XRCC2* on chromosome 7, as well as selected rodent homologs. We are also working on chromosome 19 regions associated with olfactory receptors and a congenital nephrotic disease.

We have sequenced 76 kbp containing the human and mouse *XRCC1* genes, which span 26 kbp in the mouse and 31.9 kbp in the human. In addition to the coding regions, 9 conserved elements were identified with sequence identities ranging from 65% to 78%. We have completed 54 kbp of human sequence encompassing the *ERCC2* gene as well as 54 kbp spanning the syntenic regions in the mouse and hamster. A defect in *ERCC2* leads to the cancer-prone human disorder xeroderma pigmentosum (XP-D). The human *ERCC2* gene is comprised of 23 exons and is 98% identical to the rodent homologs at the protein level. We identified two genes flanking *ERCC2*,. One may be a new member of the kinesin light chain gene family; the other has no known function. All three genes, and their orientation are conserved in the three mammals.

Like *ERCC2*, the *ERCC4* gene product is involved in the nucleotide excision repair pathway, which recognizes and removes DNA damage. The genomic region indicates that the full-length gene spans ~29 kbp and is >50% AT-rich. The *ERCC4* gene product exhibits significant homology to the *S. cerevisiae* rad1 and *S. pombe* rad16 genes, which encode single strand endonucleases.

We have sequenced a cosmid and its associated cDNA for the recently cloned human *XRCC3* gene, which appears to play a crucial role in chromosomal stability. The predicted protein shares residue identity with the GTP binding domain of the *S. cerevisiae* rad51 and rad57 proteins involved in recombinational repair. Sequence analysis of several candidate cDNAs for the XRCC2 gene also show similarity to the same domain in these proteins. Sequence analysis of the XRCC3-containing cosmid identified a second kinesin light chain gene physically linked to a DNA repair gene.

(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.)

SHOTGUN SEQUENCING OF A COSMID USING THE M13-102 DIRECT SELECTION VECTOR

Richard A. Guilfoyle, Danhua Chen, Arthur Johnson, Todd Francisco, Tetsuyoshi Ono, David Rank and Lloyd M. Smith

University of Wisconsin, Dept. of Chemistry, Madison, WI 53706

We recently reported a direct selection strategy for shotgun cloning and sequencing in the bacteriophage M13 which utilized the vector called M13-100[1]. Here, we describe the near-completed sequencing of a cosmid containing Drosophila genomic DNA using M13-102, a modified version of the vector. M13-102 contains two new additions: a homopurine•homopyrimidine tract for triple-helix-mediated affinity capture[2] (TAC), and the LacZ-derived universal primer sequences. *TAC-purification* of the linearized RF DNA is rapid - thus facilitating library construction, and specific - thus improving library quality by enabling efficient removal of non-vector DNA. *Direct selection* can (a) facilitate library production by eliminating the need for phosphatase treatment of the vector in lowering background, and (b) improve library quality by allowing phosphate treatment of the target DNA in order to reduce tandom insertion events. The cloning-efficiency analyses predicting these advantages are based on libraries constructed with a nebulizer-fragmented cosmid harboring the Drosophila GABA receptor gene. Preparation of the random single-stranded templates was performed using a streamlined solid-phase protocol developed in our laboratory and which is highly amenable to automation. Using the -21M13 dye-primers (Applied Biosystems, Inc.), fluorescent cycle-sequencing reactions were performed using Sequitherm™ (Epicentre Technologies, Madison, WI) and analyzed on ABI373A (non-stretch and stretch) machines. Based on the cosmid sequencing data, results will be presented describing the efficacy of the M13-102 strategy in terms (1) supporting the above predictions, and (2) analyses diagnostic for the production and sequencing of random clone libraries including contig assembly rates, redundancy, target coverage and gap-filling requirements.

References:
1. Guilfoyle, R.A. and Smith, L.M. (1994) *Nucleic Acids Research* **22**: 100-107.
2. Ji, H., Smith, L.M., and Guilfoyle, R.A. (1994) *GATA* **11**: 43-47

WHOLE GENOME SHOTGUN SEQUENCING OF THE 1.8 Mb GENOME OF THE HYPERTHERMOPHILIC ARCHAEA *PYROBACULUM AEROPHILUM*

Ung-Jin Kim, Steve Marsh, Ronald Swanson*, and Melvin I. Simon
Division of Biology, 147-75, California Institute of Technology, Pasadena, CA 91125.
*Recombinant Biocatalysis, Inc., Kennett Square, PA 19348.
Sorel Fitz-Gibbon and Jeffrey Miller
Department of Biology and MBI, University of California, Los Angeles, Los Angeles, CA 90024.

A random shotgun approach to sequencing approximately 1.8 Mb genome of the hyperthermophilic Archaea *Pyrobaculum aerophilum* has been undertaken. The genomic DNA of this microbe was sheared by sonication and 2-3 kb fragments were subcloned into pUC18 vectors for sequencing. The genome was also cloned in the Fosmid vector that permits stable propagation of cosmid-sized Pyrobaculum genomic DNA fragments which are unstable in high copy number cosmid vectors. The Fosmid library with 768 clones, or roughly 15X genomic coverage, was used to rapidly assemble a physical contig map of this organism using more than 100 probes as markers and restriction fingerprint analysis techniques. We are also planning to determine the end sequences of the Fosmid clones. The library and the physical map will eventually be used for gap closure at the final stage of sequence assembly.

Large numbers of sequencing templates are prepared by using an Autogen 740 miniprep machine, and are being sequenced at the Caltech sequencing core facility. Automated sequencing machines are being used to generate sequences from both ends of the template with an average of 500-600 bases of clean sequence per read. With the use of new TaqI polymeraseFS, we expect that the read length as well as the throughput will increase at least by 10-20%. As of October 1, 1995, we have obtained nearly 6,000 reads, and we expect the number will reach 25,000 or roughly 7X genomic covergage by early 1996. The sequences will then be assembled by using TIGR_ASSEMBLER and other commercially available packages, annotated, and will be deposited to publicly accessible repositories.

WHOLE GENOME SHOTGUN SEQUENCING AND ASSEMBLY OF THE *MYCOPLASMA GENITALIUM* GENOME**

Claire M. Fraser, Jeannine D. Gocayne, Owen White, Mark D. Adams, Robert D. Fleischmann, Rebecca Clayton, Kenneth F. Bott#, Hamilton O. Smith, Clyde A. Hutchison III#, and J. Craig Venter.* The Institute for Genomic Research, Gaithersburg, MD, Johns Hopkins University, Baltimore, MD* and University of North Carolina, Chapel Hill, NC#.

Mycoplasmas are small wall-less bacteria that parasitize a wide range of hosts including humans, animals, plants, insects, and tissue culture,[1,2] and are believed to represent a minimalist life form, having yielded to selective pressure to reduce genome size to eliminate unnecessary genes. *M. genitalium* currently has the smallest genome of any free-living organism (580 kb) and is believed to have evolved from ancestors common to higher gram positive organisms.[3,4] A new approach[5] for genome analysis based on whole chromosome shotgun sequencing and assembly has been successfully applied to obtain the complete nucleotide sequence (580,070 bp) of the genome of *M. genitalium*. A small insert (2 kb) *M. genitalium* genomic DNA library was constructed in pUC18 vector. A total of 5760 double-stranded DNA templates were prepared with a modified "boiling bead" method from AGTC (Gaithersburg, MD). Templates were sequenced from both ends using M13 forward (M13-21) and M13 reverse (M13RP1) primers; a total of 8472 good sequences were obtained. Assembly of the *M. genitalium* genome was accomplished with the **TIGR Assembler** which simultaneously clusters and assembles the genome. The 8472 sequences assembled into 39 contigs that ranged in size from 705 - 84,124 bp. For closure, the edited length of the sequences at the ends of the contigs was increased and the contigs were searched against each other. This process closed 11 gaps. The remaining 28 sequence gaps were closed by a single primer walk across the gaps. Analysis of putative open reading frames suggests that the *M. genitalium* genome contains approximately 470 genes. A complement of genes involved in DNA maintenance, repair, transcription, translation, and cellular transport are present; however, no pathways for amino acid, fatty acid, purine or pyrimidine biosynthesis were identified. Comparison of the *M. genitalium* genome to that of *Haemophilus influenzae*[5] suggests that differences in genome content are reflected as profound differences in physiology and metabolic capacity between these two organisms.

** Supported in part by a Department of Energy Cooperative Agreement DE-FC02-95ER61962.A000

[1]S. Razin, *Microbiol. Rev.* **49**, 419 (1985).
[2]J. Maniloff, *Mycoplasmas: molecular biology and pathogenesis*, J. Maniloff et al., Eds. (American Society for Microbiology, Washington, D.C., 1992), pp.549-559.
[3]S.D. Colman, et al., *Mol. Microbiol.* **4**, 683 (1990)
[4]C. Su and J.B. Baseman, *J. Bacteriol.* **172**, 4705 (1990).
[5]R. Fleischmann et al., *Science* **269**, 496 (1995).

COMPLETE GENOME SEQUENCING AND CHARACTERIZATION OF THE THERMOPHILIC METHANOGEN, METHANOCOCCUS JANNASCHII

Carol J. Bult, Mark Adams, Rob Fleischmann, Jeannine Gocayne, Granger Sutton, Lixin Zhou, Owen White, Lisa FitzGerald, Judy Blake, Rebecca Clayton, Ewen Kirkness, Neil Geoghagen, Jan Weidman, Joyce Fuhrmann, Brian Dougherty[1], Hanna Tomb[1], Claudia Reich[2], Claire Fraser, Gary Olsen[2], Hamilton Smith[1], Carl Woese[2], and J.C. Venter. The Institute for Genomic Research, 9712 Medical Center Drive, Rockville MD 20850; [1]Department of Molecular Biology and Genetics, The Johns Hopkins University, Baltimore MD 21205; [2]Department of Microbiology, University of Illinois, Urbana IL 61801.

The application of a whole genome shotgun strategy has been applied successfully to obtain complete genomic sequence from two Bacterial (*sensu* Woese) organisms, *Haemophilus influenzae*[1] and *Mycoplasma genitalium*[2] (URL:http://www.tigr.org). Comparison of these two genomes to each other and to sequence data from other bacteria have provided many insights into genome organization and evolution[1,2].

We have applied the whole genome shotgun strategy to sequence the genome of *Methanococcus jannaschii*. *M. jannaschii* is a barophilic, thermophilic methanogen and a member of the Archaea domain of life (*sensu* Woese). It has a genome size of approximately 2.0 Mbp and a nucleotide composition which is approximately 70% AT. Phylogenetically, *M. jannaschii* appears to be basal to other archaeal methanogens and will provide data critical for understanding the genetic basis and origin of methanogenesis.

Five aspects of the whole genome shotgun approach which are critical for its success are 1) the availability of a random genomic small insert plasmid library, 2) the availability of a representative large insert (lambda) library for the creation of a genome sequence scaffold, 3) high quality sequence data from both ends of the plasmid and lambda clones, 4) a robust sequence fragment assembly engine, and 5) data management and analysis tools which are tightly integrated with data production. For the *M. jannaschii* genome, over 36,000 sequences (~8.5 fold genome coverage) were obtained from the ends of clones from the plasmid library (average insert size 2.5 kbp). Sequence fragments were assembled using TIGR Assembler as the assembly engine[1,2,3]. Physical gaps were closed using PCR, lambda clones, and probes of genomic Southern blots. Sequences from the ends of lambda clones (average size 20 kbp) were used to build a genome scaffold to confirm the orientation and order of contigs generated by TIGR Assembler. Lambda clones also served as templates for primer walking across large repeat regions, such as the ribosomal operons.

The availability of whole genomes from both Bacteria and Archaea domains provides a solid comparative framework for describing the core genetic complement shared by these lineages, understanding what genes and biochemical pathways are unique to each domain, and testing the evolutionary relationship of Bacteria and Archaea to Eukarya (eukaryotes).

*Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy under contract DE-FC02-95ER61962.A001.

[1]Fleischmann *et al.*, *Science* **269**, 449-604 (1995).
[2]Fraser *et al.*, *Science*, in press (1995).
[3]Sutton et al., *Genome Science and Technology* **1**, 9-19 (1995).

LARGE-SCALE SEQUENCING OF THE HUMAN AND MOUSE T CELL RECEPTOR BETA LOCI

Lee Rowen, Kai Wang, Inyoul Lee, Cecilie Boysen, Ben F. Koop[1] and Leroy Hood. Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195. [1]Department of Biology, University of Victoria, Victoria, British Colombia, CA V8W 2Y2.

T cell receptors play a major role in immunity and autoimmune diseases. For this reason, their genomic sequence has been chosen as a model system for the development of strategies and tools related to the human genome project. The complete genomic sequence of a multigene locus enables a delineation of genes and gene boundaries, and an assessment of the proportion of pseudogenes. Additionally, it provides PCR access to microsatellite markers and other possible sites of polymorphic variance and, therefore, will facilitate efforts to discover mutations related to disease susceptibility. Strong sequence homologies found in a cross-species comparison between human and mouse counterparts will assist in identifying regulatory regions, new genes and alternative functions for DNA sequence information The cross-species comparison will also conduce to an understanding of the evolutionary mechanisms that underlie overall gene organization.

Over a megabase of the T cell receptor beta loci from human and mouse have been sequenced using the shotgun strategy, leading to the following discoveries.

Approximately half of the human T cell receptor beta locus is comprised of long homologous repeats in which members of multigene subfamilies are embedded. These repeats suggest a mechanism for the divergence of gene function. Indeed, a portion of the human TCR beta locus has even been translocated to another chromosome. The mouse locus, by way of contrast, contains far less repeated DNA. In this regard, the comparative genomic sequences have provided an explanation for why there are twice as many TCR beta variable gene segments in human as mouse, even though both species have about the same number of subfamilies

The T cell receptor beta locus is also the site of the human and mouse pancreatic trypsinogen multigene family, suggesting that genes with apparently unrelated functions can occupy the same genomic space. In contrast to the situation with the V betas, the mouse locus has undergone a greater expansion in the number and variety of trypsinogen genes than the human counterpart. For the most part, this expansion has not occurred through the recent duplication of DNA repeat units.

SEQUENCING OF BACS AND COSMIDS COVERING THE HUMAN T CELL RECEPTOR α/δ REGION

Cecilie Boysen, Inyoul Lee, Kai Wang, and Leroy Hood. Department of Molecular Biotechnology, University of Washington, Seattle, WA 98195.

The T cell receptors are heterodimers containing two subunits, α/β or γ/δ, recognizing specific antigenic peptides embeded in molecules encoded by the major histocompatibility complex on antigen-presenting cells. Upon recognition, the short cytosolic tails of the two antigen binding subunits transmit the signal via other T cell receptor components to the interior of the T cell, where appropriate responses occur. The T cell receptor antigen-binding chains are encoded by three different loci: α/δ, β, and γ. These loci have a very special organization of their genes in order to be able to generate large repertoires of different T cell receptors.

We have sequenced 920 kb of the α/δ locus, and anticipate to finish the last two gaps (~100 + 20 kb) later this year. Initially, we sequenced overlapping cosmids derived from YACs, but later mapped the region with BACs, which we sequenced directly applying shotgun sequencing techniques. To date, we have sequenced 5 BACs, ranging in size from 80 kb to 210 kb. These constitute the bulk of the total sequence. Phil Green in our department has developed an assembling program, phrap, which enabled us to assemble the shotgun BAC inserts without significant problems. BACs are also an attractive alternative to YACs for mapping. BAC clones are much easier to handle than YAC clones; they are rarely chimeric, and show little, if any, rearrangement. For these reasons, we have now chosen to sequence the mouse T cell receptor α/δ region using BACs.

The sequence obtained for the human T cell receptor α/δ locus has provided insights into the structure, organization, and evolution of these T cell receptor elements. It allows generation of STSs across any region, as is useful when looking for polymorphisms in connection with either genes or microsatellites. Furthermore, interesting distributions of genome-wide repeats, such as Alu or LINE sequences, have been observed over this region. Initial comparison with a ~130 kb region sequenced in mouse reveals striking sequence similarity, not only for the coding features, but also for the intergenic sequence.

The Merck Gene Index - an Update

Keith Elliston, Jeffrey Aaronson, Barbara Eckman, Richard Blevins, Shahid Imran, Joseph Myerson and Alan Williamson. Department of Bioinformatics and Research Strategies Worldwide, Merck Research Laboratories, Rahway, NJ 07065.

A key goal of the Human Genome Project is the identification of each of the estimated 100,000 genes and their location on the genome map. This will eventually be achieved by the sequencing of the complete genome and its interpretation. An alternative and immediate solution is presented by large scale partial sequencing of random cDNA clones. With this approach it is now practical to undertake a concerted effort to identify the majority of all human genes. The partial cDNA sequences generated using this approach can then be used to build the equivalent of an index to the expressed genes. Both the partial sequences and a resulting transcript map will be of extensive benefit to both academic and commercial research in the elucidation of human disease genes; a beginning point for the development of both diagnostic and therapeutic agents. Already the project has contributed over 220,000 sequences from 131,000 cDNA clones to Genbank, and many of these have been chromosomally assigned and mapped by a consortium organized by HUGO. The sequences and their corresponding cDNA clones are now available to both the academic and commercial research communities without royalty or restriction. By utilizing the sequence to organize the characterized clones into a representative set - the Merck Gene Index - which can be annotated with additional information, the project will provide a powerful resource for analysis of the human genome and genetic disease. The ongoing development of the Merck Gene Index will be discussed in detail, along with preliminary results of the analysis.

HIGH-PERFORMANCE DNA SEQUENCING USING ENERGY TRANSFER FLUORESCENT PRIMERS

Jingyue Ju,*† *Indu Kheterpal*†, *Su-Chun Hung*‡, *James R. Scherer*†, *Alexander N. Glazer*‡ *and Richard A. Mathies*†
Departments of Chemistry† and Molecular and Cell Biology‡
University of California, Berkeley, CA 94720.

Our goal is to develop novel fluorescently labeled primers for DNA sequencing and multiplex genetic analysis that have enhanced fluorescence intensities and improved match in electrophoretic mobilities.[1,2] From a library of energy transfer (ET) fluorescent primers (M13, –40) with varying donor and acceptor separation distances which we developed previously, we chose four ET primers (F10F, F10J, F10T and F10R) which satisfied the above criteria as a representative optimized set for 4-color DNA sequencing.[3] The fluorescence of these four ET primers is 2- to 14-fold greater than that of the corresponding primers labeled with only one dye when excited at 488 nm. The increased fluorescence intensity of the ET primers and the substantially similar mobilities of the DNA fragments generated with the four ET primers allow four-color DNA sequencing on a capillary array electrophoresis (CAE) DNA sequencer using a single laser line at 488 nm for excitation, and without applying mobility shift adjustments. With single-stranded M13mp18 DNA as the template, a typical run with the ET primers on a commercial ABI 373 sequencer provided DNA sequences with 99%-100% accuracy in the first 500 bases using 8-fold less DNA template (0.25 μg) than that typically required using T7 DNA polymerase. In further work, we have developed an ET cassette which can be used to label primers of any sequence, thereby significantly extending the utility of energy transfer labels.[4] We have also found that 5 or 6-carboxyrhodamine-6G can be used in ET primers to replace JOE with improved match in the mobility shift and no change in fluorescence sensitivity.[5] The design and synthesis of the ET primers and their application to high-throughput CAE DNA sequencing will be presented.

Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy under contract DE-FG-91-61125. Support from Amersham Life Science Inc. is also acknowledged.

*DOE Human Genome Distinguished Postdoctoral fellow

1. J. Ju, C. Ruan, C. W. Fuller, A. N. Glazer and R. A. Mathies. *Proc. Natl. Acad. Sci.* **92**, 4347-4351 (1995)

2. Y. Wang, J. Ju, B. A. Carpenter, J. M. Atherton, R. A. Mathies and G. F. Sensabaugh. *Anal. Chem.* **67**, 1197-1203 (1995)

3. J. Ju, I. Kheterpal, J. R. Scherer, C. Ruan, C. W. Fuller, A. N. Glazer and R. A. Mathies. *Anal. Biochem.* **231**, 131-140 (1995)

4. J. Ju, A. N. Glazer and R. A. Mathies, Universal Energy Transfer Cassette-labeled Primers for DNA Sequencing and Analysis, in preparation.

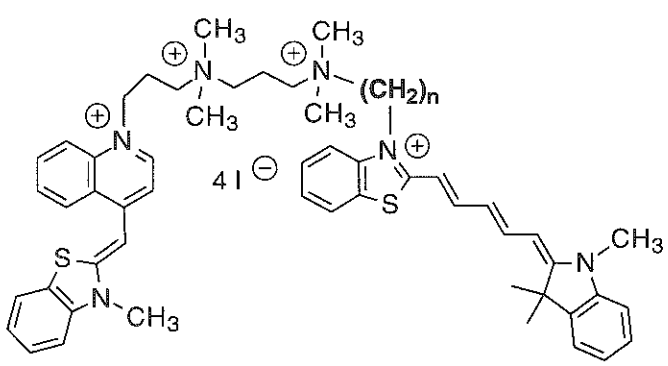
5. S-C. Hung, J. Ju, R. A. Mathies and A. N. Glazer, Energy Transfer Primers with 5 or 6-Carboxyrhodamine-6G as Acceptor Chromophores, in preparation.

FLUORESCENCE ENERGY-TRANSFER CYANINE HETERODIMERS WITH HIGH AFFINITY FOR DOUBLE-STRANDED DNA

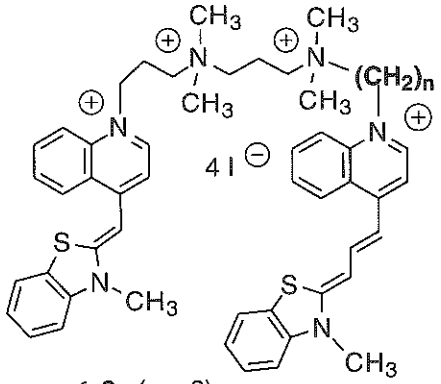
Scott C. Benson, Zhaoxian Zeng and Alexander N. Glazer, Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720.

For many applications, such as multiplex detection of double-stranded DNA (dsDNA) fragments on gel or capillary electrophoresis, two-color analyses of DNA-protein complexes, multicolor cytogenetic applications, fluorescence-activated cell sorting, and the like, it is advantageous to have sets of dyes with a common strong absorption maximum, but well-separated fluorescence emission maxima. This allows efficient excitation of several labels with readily distinguishable emission spectra with one laser line. Such sets of dyes can be generated by exploiting the phenomenon of fluorescence resonance energy transfer.[1]

We have previously reported the synthesis and spectroscopic properties of the heterodimeric intercalating cyanine dye, thiazole orange-thiazole blue (TOTAB; see structure 2a) and described its use in multiplex mapping of DNA restriction fragments[2] and in the analysis of the stoichiometry of DNA-protein complexes.[3] We have now designed, synthesized and characterized additional cyanine heterodimers that emit strongly above 650 nm with 488 nm excitation. Thiazole orange serves as the common fluorescence donor in these dyes and thiazole-indolenine (1a-c) or thiazole blue (2a-c) as acceptors.[4]



TOTIN { **1a** (n = 3); **1b** (n = 4); **1c** (n = 5) }



TOTAB { **2a** (n = 3); **2b** (n = 4); **2c** (n = 5) }

The donor emission in the dsDNA-bound dyes is quenched by over 85%. The affinity for dsDNA and the quenching of donor fluorescence is optimized by varying the length of the linker between donor and acceptor. On agarose gel electrophoresis at 10V/cm at room temperature at a basepairs:dye ratio of 20:1, the dsDNA complexes with most tightly bound dyes, 1b and 2c, gave $t_{0.5}$ values for dissociation of the dye of 317 and 1300 min, respectively.[4] Accurate two-color sizing (10 pg dsDNA per band) was obtained with 488 nm excitation with dsDNA restriction fragments precomplexed with thiazole orange dimer as unknowns (detected at 500-565 nm) and those stained with 1b or 2c (detected at 645-750 nm) as internal standards.[5,6]

*Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy under contract DE-FG-91-61125.

[1]S.C. Benson, P. Singh, and A.N. Glazer, *Nucleic Acids Res.* **21,** 5727-5735 (1993).
[2]S.C. Benson, R.A. Mathies, and A.N. Glazer, *Nucleic Acids Res.* **21,** 5720-5726 (1993).
[3]H.S. Rye, et al., *J. Biol. Chem.* **268,** 25229-25238 (1993).
[4]S.C. Benson, Z. Zeng, and A.N. Glazer, *Anal. Biochem.* **230,** (1995) in press.
[5]Z. Zeng, S.C. Benson, and A.N. Glazer, *Anal. Biochem.* **230,** (1995) in press.
[6]R.A. Mathies et al., *Rev. Sci. Instrum.* **65,** 807-812 (1994).

DNA PREPARATION AND AUTOMATION AT THE LAWRENCE BERKELEY NATIONAL LABORATORY, HUMAN GENOME CENTER

Martin Pollard

Ernest Orlando Lawrence Berkeley National Laboratory, University of California, Engineering Science Department, Human Genome Group, 1 Cyclotron Road, Berkeley, CA 94720, Phone: (510) 486-4561, Fax: (510) 486-6816, mjpollard@lbl.gov

The Lawrence Berkeley National Laboratory, Human Genome Center is sequencing Human andDrosophila DNA using a directed sequencing strategy. The Center is currently sequencing at a rate greater than 2 million base pairs per year and DNA preparation procedures are reaching a critical stage requiring automation in a 96 well format. Current progress in developing DNA preparation protocols and automation of these protocols will be described. A modified boil prep protocol is being implemented on a gantry style robot (827 W x 1715 L x 340 H mm) configured with an Eppindorf 5416 centrifuge. The robot tools consist of two 8 channel Biomek MP200 pipettors, 5 eight channel dispensing manifolds, and a pneumatic gripper for moving the microtiter plates within the work envelope and into and out of the centrifuge.

This work was supported by the Director, Office of Energy Research, Office of Health and Environmental Research, Human Genome Program, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

MODULAR, CONTINUOUS-THROUGHPUT SYSTEM FOR PROCESSING BIOCHEMICAL SAMPLES IN MICROTITER PLATES

Anthony D.A. Hansen, Martin J. Pollard, Joseph M. Jaklevic

Ernest Orlando Lawrence Berkeley National Laboratory, University of California, Engineering Science Department, Human Genome Group, 1 Cyclotron Road, Berkeley, CA 94720, Phone: (510) 486-7158, Fax: (510) 486-5857, adhansen@lbl.gov

We describe a modular conveyor-like system for the sequential processing of large numbers of samples in standard microtiter plate formats. The modules have a standard mechanical geometry to allow them to be interconnected with two parallel belts passing through to carry plates from one to the next. Each sequential module picks up the plate, performs a single task with maximum parallelism and minimum mechanical motion, and then places it back down on the moving belt. Modules may be added to or removed from the sequential lines support framing to allow for different biological protocols: modules that are installed yet not required for a particular protocol simply allow the plate to pass through. Microtiter plates are fed in to the conveyor from stacker carrying cassettes that are convenient for general use elsewhere in the laboratory: After processing the output, plates are re-stacked into empty cassettes, and waste plates are discarded. All modules include rinsing and cleaning stations for re-usable tools: no disposables are consumed.

Modules under design at the present time include:

* cassette stack plate fetcher: identification of plate type
* cassette stack plate loader
* plate filler: 48 tubes, parallel feed, 20 - 100 μl range
* reagent dispenser: 12 / 24 nozzles, individual feed, 1 - 10 μl range
* cell culture re-suspender: 96 pins in parallel
* pipet transfer station: 96 syringes, belt to belt
* UV illuminator/imaging station

To evaluate the concept in actual use our initial configuration is a plate filler, consisting of the plate fetcher, conveyor system, and multi-tube filler head. The next application will be for PCR setups from cell culture plates or tube racks.

This work was supported by the Director, Office of Energy Research, Office of Health and Environmental Research, Human Genome Program, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

Automated Fluorescent Detection for Multiplex DNA Sequencing

Andy Marks[1], *Tony Schurtz*[1], *F. Mark Ferguson*[1], *Leonard Di Sera*[1], *Alvin Kimball*[1], *Diane Dunn*[1], *Doug Adamson*[1], *Peter Cartwright*[1], *Robert B. Weiss*[1,3] *and* Raymond F. Gesteland[1,2].
[1]Department of Human Genetics and the [2]Howard Hughes Medical Institute, University of Utah, Salt Lake City, UT 84112.

Automation of a large-scale sequencing process based on instrumentation for automated DNA hybridization and detection is a focal point of our research. Recently, we have devised a method for amplifying fluorescent light output on nylon membranes by using an alkaline phosphatase-conjugated probe system combined with a fluorogenic alkaline phosphatase substrate [1]. The amplified signal allows sensitive detection of DNA hybrids in the sub-femtomole/band range.

On the basis of this detection chemistry, automated devices for detecting DNA on blotted microporous membranes using enzyme-linked fluorescence, termed Probe Chambers, have been built. The fluorescent signal is collected by a CCD camera operating in a Time Delay and Integration mode. Concentrated solutions of probes and enzymes are stored in Peltier-cooled septa sealed vials and delivered by syringe pumps residing in a gantry style pipetting robot. Fluorescence excitation is generated by a mercury arc lamp acting through a fiber optic "light line". Three 30 x 63 centimeter sequencing membranes can be simultaneously processed, currently revealing up to 108 lane sets per multiplex cycle. A probing cycle is completed approximately every eight hours.

Integration of the Probe Chamber into the production pipe line is accomplished through connections to the laboratory data base. A critical component of a high-throughput sequencing laboratory is the software for interfacing to instrumentation and managing work flow. The Informatics Group of the Utah Genome Center has designed and implemented an innovative system for automating and managing laboratory processes. This software allows the model of workflow to be easily defined. Given such a model, the system allows the user to direct and track the flow of laboratory information. The core of the system is a generic, client-server process management engine that allows users to define new processes without the need for custom programming. Based on these definitions, the software will then route information to the next process, track the progress of each task, perform any automated operations, and provide reports on these processes. To further increase the usefulness of our laboratory information system, we have augmented it with hand-help mobile computing devices (Apple Newtons) that link to the database through RF networking cards.

Base calling software has been developed to support our automated, large scale sequencing effort. 1st stage sequence calling identifies putative bands, however, depending on the number of reader indel errors (2-6%), merging 1st stage sequence without the aide of cutoff information can be difficult. To improve our base calling we have employed Fuzzy Logic to establish confidence metrics. The logic produces a confidence metric for each band using band height, width, uniqueness, shape, and the gaps to adjacent bands. The confidence metric is then used to identify the largest block of highest quality sequence to be merged.

This work was funded, in part, by DOE grant DE-FG03-94ER-61817 (R.F.Gesteland, P.I.)

[1] Cherry, J.L., Young, H., Di Sera, L.J., Ferguson, F.M., Kimball, A.W., Dunn, D.M., Gesteland, R.F., and Weiss, R.B. (1994). Enzyme-linked fluorescent detection for automated multiplex DNA sequencing. *Genomics* **20**, 68-74

DNA Sequencing by Single Molecule Detection

R.L. Affleck, J.N. Demas, P.M. Goodwin, J.H. Jett, R.A. Keller,
J.C. Martin, J.A. Schecker, D.J. Semin, and M. Wu

Center for Human Genome Studies
Los Alamos National Laboratory

ABSTRACT

We are developing a technique to determine the sequence of bases in large fragments of DNA. Our goal is a sequencing rate of 100 to 1000 bases per second on DNA strands approaching 40 kb in length. The ideas presented represent the combined effort of a multidisciplinary team composed of physicists, physical chemists, cellular and molecular biologists and organic chemists. A large fragment of DNA, approximately 40 kb in length, will be labeled with base identifying fluorescent tags and suspended in the flow stream of a flow cytometer capable of single fluorescent molecule detection. The tagged bases will be cleaved sequentially from the single DNA fragment and identified by laser-induced fluorescence as they pass through the excitation laser beam. We have demonstrated that each of the component parts of the technology work separately and have integrated them into a prototype instrument. Using this prototype, we have demonstrated recently the digestion of fluorescently labeled DNA fragments anchored in a flowing stream containing E. coli exonuclease III. We have detected single labeled nucleotides, enzymatically cleaved from 15-30 DNA fragments anchored in the flow stream. This is a major milestone towards the realization of DNA sequencing in flow. From this data we estimate the exonuclease turnover rate in this system to be $\sim$5 nucleotides per second per DNA strand at 36° C, similar to the rate measured under static conditions for this exonuclease and native DNA. We are exploring different ways to increase the exonuclease turnover rate to $\sim$100 nucleotides per second per DNA strand.

High Efficiency, Single Molecule Detection for DNA Sequencing

Rhett L. Affleck†, James N. Demas†, Peter M. Goodwin†,
James H. Jett††, Ming Wu†, and Richard A. Keller†

† Chemical Science and Technology Division
†† Life Sciences Division

Los Alamos National Laboratory
Los Alamos, NM 87545

The Los Alamos, flow cytometry based approach to DNA sequencing requires highly efficient single molecule detection. We have demonstrated a detection efficiency approaching 95% for single molecules delivered from a small source in the center of our flow cell. The efficiency for single molecule detection is limited by false positives from fluorescent impurities in buffers and enzyme solutions, radial diffusion of analyte molecules out of the center of the sample stream, and photobleaching of analyte molecules during their transit through the excitation laser beam.

We have also demonstrated that: (1) in-line photobleaching of fluorescent impurities present in the enzyme and buffer solutions reduces our background significantly and (2) a high molecular weight polymer added to the sheath stream forms complexes with the small analyte molecules in the sample stream resulting in reduced radial diffusion of the fluorescent adduct.

Details of these measurements and applications to our DNA sequencing project will be presented.

This work was supported by Los Alamos National Laboratory LDRD funds and the DOE/OHER Human Genome Program.

ONE-STEP PCR SEQUENCING*

Kenneth W. Porter, J. David Briley, and Barbara Ramsay Shaw

Department of Chemistry, Duke University, Durham, NC 27708

A method to amplify and sequence DNA simultaneously by incorporating potential chain delimiters is described. During PCR amplification, a small percentage of boron modified nucleotides (*e.g.*, 2'-deoxynucleoside 5'-α-P-borano-triphosphates[1,2]) are incorporated into the product DNA. The positions of the boranophosphates can be revealed by exonuclease digestion, thereby defining the sequence of the PCR product. The One-Step method improves current PCR sequencing methods by avoiding both DNA purification following amplification and single-sided primer extension with dideoxynucleotide chain terminators. As a consequence, One-Step sequencing is fast and amenable to automation. Data obtained by the One-Step method is comparable to that produced by cycle sequencing, yet requires much less DNA template.

A region of the p53 gene was sequenced by the One-Step boranophosphate method. The sequencing primer was modified with Cy5 and the resultant sequencing fragments were analyzed by an automatic fluorescent sequencer. As a comparison, the same region was sequenced by conventional cycle sequencing. For both the One-Step and cycle sequencing methods, the sequence could be read over approximately 450 bases. Sequence quality was similar for each method. In each case where a particular base could not be determined by one method, the base could often be called correctly by the other method. Therefore the One-Step method produces data that is comparable, as well as complementary, to dideoxy sequencing.

One-Step Sequencing should offer numerous advantages for DNA sequencing:

- speed because sequence delimiters are incorporated during PCR
- accuracy because both strands can be sequenced easily
- ease of automation because template purification is unnecessary
- convenience because only minute quantities of template are required.

* Supported by NIH Grant HG00782 and DOE Grant DE-FG05-94ER61882.

1 Sood, A., Shaw, B. R., and Spielvogel, B. F. (1990) *J. Amer. Chem. Soc. 112*, 9000-9001.

2 Tomasz, J., Shaw, B. R., Porter, K., Spielvogel, B. F., and Sood, A. (1992) *Angew. Chem. Int. Ed. Engl. 31*, 1373-1375.

THE COW DNA SEQUENCER - INSTRUMENTATION AND SOFTWARE FOR UNATTENDED COSMID ORIENTED WALKING AND LARGE SCALE DNA SEQUENCING OF THE HUMAN GENOME.

Harold R. Garner, David Burbee, Stafford Brignac, Kim Burzynski, Christopher Davies, Michelle Gilbert, Ken Kupfer,[1] *Ping Li, Shane Probst, Simon Rayner, Emilee Strunk and Glen A. Evans,* McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center at Dallas, 6000 Harry Hines Boulevard, Dallas Texas, 75235-8591.

Cosmid oriented walking ("COW") is a novel and potentially powerful approach for determining the sequencing of human chromosomes based on sampled maps at any required level of accuracy. COW sequencing depends on the simultaneous sequencing of cosmid clones and the construction of contigs based on sequence matching to a large cosmid-end-sequence database. One advantage of COW sequencing is the potential for complete, unattended automation. A COW DNA sequencer prototype is being constructed which will carry out dye terminator DNA sequencing from 96 cosmid templates simultaneously, processing of 96 sequences, prediction of oligonucleotide "walking" primers for extending the sequence of each fragment, programming a network attached 96-channel oligonucleotide synthesizer, synthesis of the primers and placement of the primers into the cosmid template set to initiate a second round of sequencing. Using a set of nested cosmids covering 800 kb at 5X redundancy, COW sequencing should allow completion of 800 kb in 8 to 16 cycles. Such an instrument for unattended DNA sequencing is currently under construction.

The enabling technologies for COW sequencing are high quality dye terminator chemistry based sequencing on cosmids, the implementation of a high-throughput, low cost automatic oligo synthesizer and development of automated control and sequence assembly software. An oligo-synthesizer based on a 96 channel upgrade to the LBL 12-channel synthesizer, but using Macintosh hardware and Labview control software is being assembled. Network software based on Applescript and script-aware software for parallel assembly of 96+ independent walks is being written using a combination of commercial and in-house developed software packages.

Supported by a grant from the Office of Health and Environmental Research, Department of Energy.
[1]SERCA Fellow of the National Center for Human Genome Research
Supported in part by the National Center for Human Genome Research, the Whitaker Foundation and the Eugene McDermott Foundation.

COW (COSMID ORIENTED WALKING) DNA SEQUENCING: A NOVEL STRATEGY FOR RAPID, DIRECT, AND AUTOMATED SEQUENCING OF THE HUMAN GENOME BASED ON SEQUENCE SAMPLED MAPS.

Dave Burbee, Chris Davies, Michelle Gilbert, Kim Burzynski, Emilee Strunk, Jeff Schageman, James McFarland, Ken Kupfer[1], Skip Garner, and Glen A. Evans.

Recent developments in DNA sequencing technology allow the rapid analysis of multiple DNA samples. We have recently shown that with special template preparation and with the use of dye terminator chemistry, large DNA molecules such as intact cosmids can be readily analyzed. We are able to collect 400 to 600 base runs with > 98% accuracy of cosmid templates. For random sets of oligonucleotide primers, greater than 90% efficiently prime the Taq FS-catalyzed sequencing reaction. Also, recent developments in oligonucleotide synthesis have resulted in the decreased cost of custom-designed oligonucleotides. We are able to rapidly and inexpensively sequence cosmids by standard walk procedures: T7 and T3 promoter primers are used as entry points to sequence the insert, and each successive round of sequencing is used to design the next oligo that primes the sequencing reaction farther within the interior of the insert.

Over the past few years, a large number of strategies have been designed for the sequencing of large genomes, including standard shotgun sequencing, sequencing of nested deletions, and transposon-mediated primer insertion (dog-tagging). We propose a new strategy that extends our current development and use of GSS (genome sequence-sampled) maps, and is based on our success with cosmid template sequencing. We have completed a 100 kb resolution map of human chromosome 11 based on standard STS's as well as STS's designed from the sequencing of 17000 ends of a chromosome 11-specific cosmid library. Many of these cosmids have been associated by mapping of overlapping restriction fragments and/or binned within specific YACs by IRE-bubble PCR hybridization. This database was generated to identify start points for the cosmid-oriented walking (COW) DNA sequencing strategy. COW sequencing is initiated by the complete sequencing of a cosmid or cosmid contig by primer-mediated walking or else shotgunning. Extension of the initial contig is carried out by searching the GSS database for other identified cosmid ends, and so extend the sequence with the sequencing of the contiguous overlapping cosmid. Using this technique, cosmid template sets extending over several megabases may be completed.

Though ultimately limited by the cost of oligonucleotide primers, primer walking has several advantages over the current strategies. Little redundancy of sequence template preparation and analysis is required; specific regions of interest may be identified (by GSS) and then analyzed rapidly and economically. Furthermore, coupling of automated DNA sequencing instrumentation to DNA sequence analysis programs and multichannel oligonucleotide synthesizers will allow almost complete automation of the cosmid oriented walking strategy, allowing a small group of researchers to produce (for a 96 channel oligosynthesizer) 40 to 60 kb of DNA sequence per round of sequencing.

Supported by a grant from the Office of Health and Environmental Research, Department of Energy.

[1]SERCA Fellow of the National Center for Human Genome Research

Supported in part by the National Center for Human Genome Research, the Whitaker Foundation and the Eugene McDermott Foundation.

INTEGRATION OF LABORATORY AUTOMATION FOR THE HUMAN GENOME PROJECT: THE SEQUATRON.

Trevor L. Hawkins., Cheryl Evans., David Goon., Jarrod Loncor., Tara O'Connor., Jim McDermott., Niall Moloney., Meghan Lane., William Fitzhugh., David Wang., and William Lee. Whitehead Institute/ MIT, Center for Genome Research, One Kendall Square, Cambridge, MA 02139.

There is a major need for the development of a system which can accomplish the integrated tasks of DNA isolation and proceed with purification and the setup of sequencing reactions. Here we demonstrate the feasibility of such a system from both a biochemical and engineering perspective.

We have set up collaborations with CRS robotics corp, Packard Instruments, Tecan US and Techne Inc. to design and construct a 'factory style' laboratory system which is called the Sequatron..

The major component of our system is an articulated CRS 255A robotic arm which is track mounted. The deck of the robot contains several new/modified XYZ robotic workstations, a novel thermal cycler with automated headed lids, carousels and custom built plate feeders.

Biochemically, we have employed our Solid-phase reversible immobilization (SPRI) technique to isolate and manipulate the DNA throughput the process. Specifically we have set up the Sequatron to isolate DNA from M13 phage or crude PCR products using the same protocol and procedures. From M13 phage we obtain approximately 1µg of DNA per well which is sufficient for multiple sequencing reactions.

The current throughput of the system is 80 microtiter plates of samples from M13 phage supernatants or crude PCR products to sequence ready samples every 24 hours. Recently, new enzymes, new energy transfer primers and higher density microtiter plates have opened up possible increases to in excess of 25,000 samples per 24 hour period.

Supported by a grant from the Office of Health and Environmental Research of the U.S. Department of Energy under contract DE-FG02-95ER62099.

DNA Sequencing Technology: Walking with Modular Primers

Levy Ulanovsky, Lev Kotler, Mugasimangalam Raja, Maya Shmulevitz and Alexander Beskin
Dept. of Structural Biology, Weizmann Institute of Science, Rehovot, 76100 Israel.
Phone: +972-8-343547; FAX: +972-8-344105
E-mail: BPLEVY@WEIZMANN.WEIZMANN.AC.IL

We are developing DNA sequencing technology using modular primers, which eliminate the primer synthesis, the main bottleneck in primer walking. Modular primers are assembled from three 5-mer, 6-mer or 7-mer modules selected from a presynthesized library of as few as 1000 oligonucleotides. The three modules anneal contiguously at the selected template site and prime there uniquely, even though each is not unique for the most part when used alone. This technique is expected to speed up primer walking 20 to 50 fold, and reduce the sequencing cost by a factor of 5 to 15. Time and expense will be saved not only on primer synthesis itself but, more important, because the instant availability of the primers enables closed-loop automation of the complete cycle of walking sequencing. Apart from saving time and cost, the closed-loop automation would largely eliminate the need for human intervention between the walks, as a source of errors and complications.

We have found that modular primers can be used with dye-terminators of the ABI 373A automated sequencer. Also, reactions with the modular primers and dye terminators were run successfully on replaceable matrix capillaries of Barry Karger (Northeastern University) in 60 min. and read beyond 500 bases with as few as 2 base-calling errors. Because no protein needs to be removed, no precipitation or phenol extraction (obstacle to closed-end automation) is required. The success rate and quality of automated sequencing with modular primers are similar to those with conventional primers 17-20 base long. For the most part few, if any, base-calling errors are found within the first 400 bases of the sequence run. We currently prefer pentamer-based primers of the 5+7+7 structure with Pu-Pu base stacking between the 5-mer (to be extended) and adjacent 7-mer. Both heptamers are 3'-end modified to prevent their extension by the polymerase; each has two degenerate positions and thus the same size library as the pentamers (512 sequences).

*Supported by DOE grant De-FG02-94ER61831.

SIDE EXCITATION OF FLUORESCENCE IN ULTRATHIN SLAB GEL ELECTROPHORESIS†

Danhua Chen, Mark D. Peterson, Robert L. Brumley, Jr., Michael C. Giddings, Eric C. Buxton, Michael S. Westphall, Lloyd Smith and Lloyd M. Smith*, Department of Chemistry, University of Wisconsin-Madison, Madison WI 53706, *Lawrence Berkeley Laboratory, Berkeley CA.

Electrophoresis in thin gels provides increased heat transfer efficiency, permitting larger electric fields to be employed with correspondingly more rapid separations[1,2]. This is of particular interest in the area of fluorescence-based automated DNA sequence analysis, where there is a tremendous need for increased throughput from sequencing instruments[3]. Utilizing a four color cooled CCD camera in conjunction with ultrathin slab gels, Kostichka et al. demonstrated an order-of-magnitude increase in separation speed for fluorescence-based DNA sequencing[4].

This prototype apparatus, which had a narrow field of view, was optimized to include a larger CCD chip in the detection system. This allowed a factor of three increase in the imaged area, thus allowing additional samples to be run in parallel. Sample excitation for this the system was accomplished by bringing the laser beam into the gel from the side. This method of excitation permitted the required excitation power density to be obtained with an air cooled argon ion laser, as the excitation beam cross section remains small. The amount of background fluorescence is also reduced since the beam does not pass through the glass used to cast the sequencing gel.

This approach has been used successfully for conventional sequencing gels about 400 microns in thickness, and is employed in commercial sequencing instruments from Hitachi and Pharmacia. However, the fundamental properties of gaussian laser beams introduce problems when trying to pass the beam through an ultrathin gel. Namely, the tighter the focus of the beam the shorter the distance over which the focus can be maintained. Fortunately, the high efficiency of grazing incidence reflection effectively traps the beam between the glass plates, resulting in a high throughput of the laser energy.

A theoretical model describing the beam throughput has been developed. In this model attenuation of the beam intensity is attributed to four factors: aperturing at the entrance of the gel; reflective losses upon entrance into the gel; scattering during transmission through the gel; and reflective losses occurring upon successive "bounces" of the beam from the gel-glass interface during propagation of the beam. The beam properties as characterized theoretically are shown to be in good agreement with the experimentally determined values.

†Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy under contract DE-FG02-90ER61026

[1]Drossman, H.; Luckey, J. A.; Kostichka, A. J.; D'Cunha, J.; Smith, L. M. *Anal. Chem.* **1990**, *62*, 900-903.

[2]Brumley, R. L.; Smith, L. M. *Nucl. Acids Res.* **1991**, *19*, 4121-2126.

[3]Hunkapiller, T.; Kaiser, R. J.; Koop, B. F.; Hood, L. *Science.* **1991**, *254*, 59-67.

[4]Kostichka, A. J.; Marchbanks, M. L.; Brumley, R. L.; Drossman, H.; Smith, L. M. *Bio/Technology.* **1992**, *10*, 78-81.

DNA SEQUENCING AND ANALYSIS BY MULTI-CAPILLARY ELECTROPHORESIS USING REPLACEABLE LINEAR POLYACRYLAMIDE*

Barry L. Karger, Emanuel Carrilho, Jan Berka, Marie Ruiz-Martinez[1], Frantisek Foret, Steve Carson and Arthur Miller, Barnett Institute and Department of Chemistry, Northeastern University, Boston, MA 02115.

The goal of this research is to construct a robust, general purpose multi-capillary instrument for high throughput DNA sequencing and analysis using replaceable polymer matrices. The instrument design is based on no moving parts and full illumination in the detection region in order to operate each column with the fastest speeds. The laser light is split into individual beams for constant irradiation of the column. The emitted fluorescence light is collected by a wide angle lens and thence on a CCD for specific dye determination. A significant effort has been conducted on column design. For example, highly stable hydrophilic wall coatings have been developed for long term column use. Secondly, the linear polyacrylamide polymer solution has been optimized in terms of polymer molecular weight and concentration for sequencing at least 500 - 600 bases in roughly a one-hour run. Furthermore, procedures have been developed for highly reproducible large scale polymerization procedures, necessary for rugged operation. The role of column temperature has also been carefully examined to increase separation speed and reduce band compressions. These studies have been coupled with appropriate sample clean-up procedures for multi-capillary operation. In addition, base-calling software routines are being developed based on digital communication procedures for sequence readings, including numerical estimates of base confidences. We will report on our latest advances, for a variety of sequencing strategies, including primer walking with short oligonucleotide primers.

*Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Resarch of the U.S. Department of Energy under contract DE-FG-90-60985.

[1]DOE Human Genome Distinguished Postdoctoral Fellow.

Multiple Capillary DNA Sequencer that uses Fiber-Optic Illumination and Detection

Mark A. Quesada, Shiping Zhang, Janine Graves and F. William Studier
Biology Department, Brookhaven National Laboratory, Upton, New York 11973

An 8-capillary prototype electrophoresis system for DNA sequencing has been constructed. The sequence of more than 400 bases can be obtained from each capillary in less than an hour, from sequencing reactions generated with ABI four-color fluorescent terminators. Illumination of each capillary and collection of fluorescence is through individual optical fibers. Resolution of the DNA ladder is through a replaceable sieving matrix of linear polyacrylamide in re-usable coated capillaries generously provided by Karger's group (1).

Light from an argon ion laser is introduced into a fused biconically tapered fiber-optic splitter, and individual fibers deliver approximately 10 mW of 514 nm light to each of the 8 electrophoresis capillaries. Illumination and collection are by fibers normal to the surface of the electrophoresis capillary and at right angle to each other. Illumination by a fiber with low numerical aperture and collection by a fiber with high numerical aperture provides good sensitivity and signal-to-noise ratios without the need for micro-lenses. The OH stretch Raman line was convenient for aligning the fibers. In the prototype version, the fibers are fixed to the electrophoresis capillary with optical cement, but more convenient configurations are being explored. The 8 collection fibers are passed in parallel through holographic filters for Rayleigh rejection and into an imaging spectrograph, which simultaneously displays the full fluorescence spectrum (500-670 nm) from the 8 capillaries in parallel on the surface of an intensified CCD. The CCD is read out at a rate of 3.4 complete images per second. We are developing base-calling software for this system and intend to generate confidence levels as an intrinsic part of the base-calling process.

Improvements of this prototype system will be aimed at developing a reliable and fully automated system capable of sustained production of tens of kb of sequence per hour.

Supported by the Office of Health and Environmental Research of the U. S. Department of Energy.

(1) Ruiz-Martinez, M. C., Berka, J., Belenkii, A., Foret, F., Miller, A. W., and Karger, B. L. DNA sequencing by capillary electrophoresis with replaceable linear polyacrylamide and laser-induced fluorescence detection. *Anal. Chem.* **65**, 2851-2858 (1993).

DEVELOPMENT OF INSTRUMENTATION FOR DNA SEQUENCING AT A RATE OF 40 MILLION BASES PER DAY

Edward S. Yeung, Huan-Tsung Chang, Qingbo Li, Xiandan Lu, Eliza Fung, Ames Laboratory-USDOE and Department of Chemistry, Iowa State University, Ames, IA 50011.

We have developed novel separation, detection, and imaging techniques for real-time monitoring in capillary electrophoresis. These techniques will be used to substantially increase the speed, throughput, reliability, and sensitivity in DNA sequencing applications in highly multiplexed capillary arrays. We estimate that it should be possible to eventually achieve a raw sequencing rate of 40 million bases per day in one instrument based on the standard Sanger protocol. We have reached a stage where an actual sequencing instrument with 100 capillaries can be built to replace the Applied Biosystems 373 or 377 instruments, with a net gain in speed and throughput of 100-fold and 24-fold, respectively.

The substantial increase in sequencing rate is a result of several technical advances in our laboratory. (1) The use of commercial linear polymers for sieving allows replaceable yet reproducible matrices to be prepared that have lower viscosity (thus faster migration rates) compared to polyacrylamide. (2) The use of a charge-injection device camera allows random data acquisition to decrease data storage and data transfer time. (3) The use of distinct excitation wavelengths and cut-off emission filters allows maximum light throughput for efficient excitation and sensitive detection employing the standard 4-dye coding. (4) The use of index-matching and 1:1 imaging reduces stray light without sacrificing the convenience of on-column detection.

Continuing efforts include further optimization of the separation matrix, development of new column conditioning protocols, refinement of the excitation/emission optics, design of a pressure injection system for 96-well titer plates, validation of a new 2-color base-calling scheme, simplification of software to allow essentially real-time data processing, implementation of voltage programming to shorten the total run times, and scale up of the technology to allow parallel sequencing in up to 1,000 capillaries.

K. Ueno and E. S. Yeung, "Simultaneous Monitoring of DNA Fragments Separated by Capillary Electrophoresis in a Multiplexed Array of 100 Channels", Anal. Chem. **66**, 1424-1431 (1994).

X. Lu and E. S. Yeung, "Optimization of Excitation and Detection Geometry for Multiplexed Capillary Array Electrophoresis of DNA Fragments", Appl. Spectrosc. **49**, 605-609 (1995).

Q. Li and E. S. Yeung, "Evaluation of the Potential of a Charge Injection Device for DNA Sequencing by Multiplexed Capillary Electrophoresis", Appl. Spectrosc. **49**, 825-833 (1995).

E. N. Fung and E. S. Yeung, "High-Speed DNA Sequencing by Using Mixed Poly(ethyleneoxide) Solutions in Uncoated Capillary Columns," Anal. Chem. **67**, 1913-1919 (1995).

Q. Li and E. S. Yeung, "Simple Two-Color Base-Calling Schemes for DNA Sequencing Based on Standard 4-Label Sanger Chemistry", Appl. Spectrosc. **49**, 1528-1533 (1995).

MULTIPLE CAPILLARY DNA SEQUENCING

Norman J. Dovichi, Jian-Zhong Zhang, JuYing Yan, Jiang Rong, Rong Liu, Sue Bay, Pieter Roos, Karl Voss, Scott Dellinger

Department of Chemistry, University of Alberta, Edmonton, Alberta CANADA T6G 2G2

We have developed multiple capillary DNA sequencers. These instruments have several important attributes. First, by operation at electric fields greater than 100 V/cm, we are able to separate DNA sequencing fragments rapidly and efficiently. Second, the separation is performed with 3%T 0%C polyacrylamide. This low viscosity, non-crosslinked matrix can be pumped from the capillary and replaced with fresh material when required. Third, we operate the capillary at elevated temperature. High temperature operation eliminates compressions, speeds the separation, and increases the read length. Fourth, our fluorescence detection cuvette is manufactured locally by means of microlithography technology. These detection cuvettes provide robust and precise alignment of the optical system. Currently, 5, 16, and 90 capillary instruments are in operation in our lab; 32 and 576 capillary devices are under development. Fourth, we use both avalanche photodiode photodetectors and CCD cameras for high sensitivity detection. We have obtained detection limits of 120 fluorescein molecules injected onto the capillaries. High sensitivity is important in detecting the low concentration fragments generated in long sequencing reads. This combination of low concentration acrylamide, high temperature operation, and high sensitivity detection allows separation of fragments over 800 bases in length in 90 minutes.

DEVELOPMENT OF A HIGH-THROUGHPUT AND HIGH-SENSITIVITY CAPILLARY ARRAY DNA SEQUENCER

Jian Jin, William F. Kolbe and Jocelyn C. Schultz

Ernest Orlando Lawrence Berkeley National Laboratory, University of California, Engineering Science Department, Human Genome Group, 1 Cyclotron Road, Berkeley, CA 94720, Phone: (510) 486-4082, Fax: (510) 486-5857, jian_jin@lbl.gov

It is likely that in the next generation of DNA sequencing machines, capillary array electrophoresis will be a major technology for scale-up for the Human Genome Project. In order to provide high-speed, high-throughput and cost-effective DNA sequencers to the sequencing team at LBNL's Human Genome Center and elsewhere, we have recently initiated a project to develop a practical and useful capillary array DNA sequencer. The goal of this project is to develop a fully automated DNA sequencing machine, having 96 capillary columns using a replaceable sieving matrix. The design philosophy of this machine will be to incorporate existing research results in fluorescence capillary electrophoresis studies, and implement the automation of the capillary coating, gel-filling/replacing and sample injection. To this end, we have developed a 24-capillary prototype DNA sequencing machine. The system employs the sheath-flow technique developed by N. J. Dovichi's group [1], in which an excitation laser beam traverses a gel-free flow cell and irradiates all DNA migration lanes, resulting in high-sensitivity DNA fragment detection. The fluorescence image is recorded by a cooled CCD camera. Using 6% linear polyacrylamide gel as sieving matrix, we are able to separate fragments at a speed of 500 bases/hour/capillary with an electric field of 300 V/cm, and obtain single base resolution up to 400 bases. The system is constructed in such a way that a gel-filled array of capillaries can be directly plugged into the flow cell and run immediately. After each run, the capillary array is unplugged from the cell and sent to a gel-refilling station. We have also built a sample injection apparatus which requires less than 1 μl of sample solution per capillary, and permits simultaneous sample injection for the whole array. Currently a 24-capillary, four color machine, which employs an image-splitting prism to separate the four colors, is being tested. The final test results will be presented.

[1] Cheng Y.F., Wu S., Chen D.Y., Dovichi N.J., Anal. Chem., 62, 496-503 (1990).

This work was supported by the Director, Office of Energy Research, Office of Health and Environmental Research, Human Genome Program, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

HIGH-THROUGHPUT FLUORESCENT SEQUENCER DEVELOPMENT

William F. Kolbe, Jocelyn C. Schultz and Jian Jin

Ernest Orlando Lawrence Berkeley National Laboratory, University of California, Engineering Science Department, Human Genome Group, 1 Cyclotron Road, Berkeley, CA 94720, Phone: (510) 486-7199, Fax: (510) 486-5857, wfkolbe@lbl.gov

In order to address the need for increased sequencing capability, we have undertaken a program of gel-based fluorescent sequencer development. Currently available slab-gel fluorescent sequencers are limited in the number of sample lanes they can accommodate due to limitations in the spatial resolution of their detection systems. We have developed a fiber optic based detection system permitting at least 100 lanes to be used in a 25 cm width. This detection system has been applied to two different gel platforms: a conventional sized (30 cm x 25 cm x 350 μm) and an ultra-thin gel apparatus (15 cm x 25 cm x 100 μm).

The detection system uses a laser beam passing transversely through the gel to excite all the DNA lanes simultaneously and a fiber optic array to collect the fluorescence produced. A gradient index lens array is employed to image the fluorescence on to the ends of the fibers. The output end of the fiber array is formed into a compact rectangular shape compatible with a cooled charge-coupled device camera used to detect the light. Both single color and four color operation have been investigated. To provide four color detection capability, a motorized translation stage containing interference filters is placed between the camera and the fiber optic array.

Because of the high spatial resolution of the optical system, alignment of the laser beam and detector array is critical. In addition, the laser beam is found to interact with the gel material eventually producing a damage zone which causes a deflection and subsequent broadening of the beam. To eliminate these problems, the alignment of the laser beam is stabilized by means of a feedback system employing a motorized steering mirror and beam translation stage controlled by photodiode detectors positioned at each end of the fiber optic array. To eliminate gel damage by the laser beam, the detector array assembly together with the stabilized laser beam is slowly scanned over the course of a sequencing run. The scanning rate employed (0.5 mm per hour) is sufficiently slow that no observable effect on the quality of the sequencing data is produced.

The two systems were characterized using sequence data generated with M13mp18 templates. The performance was evaluated on the basis of resolution, speed and sequence accuracy using in-house base calling software.

This work was supported by the Director, Office of Energy Research, Office of Health and Environmental Research, Human Genome Program, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

DEVELOPMENTS IN HIGH THROUGHPUT ELECTROPHORESIS SYSTEM FOR DNA SEQUENCING AND LARGE FRAGMENT ANALYSIS

Joseph W. Balch, Courtney Davidson, Larry Brewer, Jackson Koo, Ray Mariella, and Anthony Carrano. Human Genome Center, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94550

We have been investigating the design alternatives and fabrication technologies necessary for building highly integrated, high throughput DNA sequencing and large fragment analysis systems based on electrophoresis. Our ultimate objective is to develop a DNA analysis system that will sequence up to 3000 clones per 24 hours of operation. Our preliminary design for such an ultimate sequencer uses 384 lanes per system and is capable of resolving greater than 500 bases per lane. We are using microfabrication techniques to build arrays of electrophoresis microchannels on glass substrates. This is an alternative technology[1] to bundles of discrete capillaries that are being investigated by others[2,3,4,5,6]. We believe the microfabrication approach will allow the assembly of a more physically robust system and provide the foundation for ultimately integrating chemical micro-reaction chambers and other fluidic hardware to allow more automated processing of DNA samples for sequencing.

In 1994 we reported the fabrication of 48 channels (1 mm wide, 0.2 mm deep and 25 cm long) etched in large glass plates (25 cm x 42 cm) that resolved about 500 bases per channel[7]. These were used in a 'hybrid' mode where adjacent channels were not physically sealed but separated by a thin layer (~50 μm) of gel. To significantly increase the density of microchannels on a practical size glass plate we have had to develop sealed microchannels to prevent "cross-talk" of DNA among adjacent channels. We have bonded glass plates with etched microchannels which are 10 cm x 10 cm. Realizing that longer channels are desirable to obtain long base reads (e.g. greater than 500 bases), we have also built and are testing the performance of alternative geometries for separation columns (e.g. serpentine loops vs. straight channels). Serpentine or looped channels can be used to arbitrarily extend the overall length of channels at the cost of packing density of channels and some loss of resolution depending upon the number of turns per column[8]. We will discuss experimental results as well as electric field modeling results which have suggested modifications of the geometry of the bends can reduce band spreading at column turns. The trade off of increased channel length necessary for extended base read and the number of channels per substrate has led to a continuing effort to develop a viable bonding process for larger glass plates. In an effort to meet the requirement of extended base read while considering practical limitations on large area bonding we have also developed a model of electrophoretic resolution to help us better understand and optimize the design of our separation channels. We will report on results obtained using this model and discuss the attendant ramifications for the overall system design.

(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.)

---

[1] A.T. Wooley and R. A. Mathies, Proc. Natl. Acad. Sci. USA, Vol 91 pg 11348 (1994)
[2] X. C. Huang, M. A. Quesada, and R. A. Mathies, Anal. Chem., Vol. 64, pg 2149 (1992)
[3] K. Ueno and E. S. Yeung, Anal. Chem., Vol 66, pg 1424 (1994)
[4] N. Dovichi, 4th DOE Human Genome Program Contractor-Grantee Workshop Santa Fe, NM (1995)
[5] B. L. Karger et. al., 4th DOE Human Genome Program Contractor-Grantee Workshop Santa Fe, NM (1995)
[6] M. A. Quesada et al., 4th DOE Human Genome Program Contractor-Grantee Workshop Santa Fe, NM (1995)
[7] J. Balch, C. Davidson, et al, International DNA Sequencing Conference, Hilton Head N. C., (1994)
[8] S. C. Jacobson, R. Hergenroder, et. al., Anal. Chem., Vol. 66, pg 1107 (1994)

THIRD-GENERATION DNA SEQUENCING AND ANALYSIS TECHNOLOGIES: CAPILLARY ARRAY ELECTROPHORESIS CHIPS AND SINGLE MOLECULE DNA FRAGMENT DETECTION

Richard A. Mathies, Steve M. Clark, Andrew J. de Mello, Brian B. Haab, Jingyue Ju, Indu Kheterpal, James R. Scherer, Yiwen Wang and Adam T. Woolley
Department of Chemistry, University of California, Berkeley CA 94720

This presentation will focus on the portion of our work concerned with the development of miniaturized capillary array electrophoresis (CAE) chips and integrated DNA analysis devices as well as the practical extension of detection sensitivity to the single molecule limit.

CAE Chips. To reduce the electrophoretic lane dimensions and to increase separation speed we have developed microfabricated capillary arrays using photolithographic masking and chemical etching techniques. High resolution restriction and PCR fragment sizing separations can be performed on these chips in under 120 s. Using polymerized gels, these chips can separate DNA sequencing reactions out to 400 bases in under 10 minutes. Raw sequencing rates for a single microfabricated capillary are ~1000-2000 bases/hour. We are currently working with a 32 capillary chip that can perform simultaneous separations of 32 different samples in parallel. We are also working on integrating PCR sample preparation on these chips to produce miniaturized integrated DNA analysis systems.

Single Molecule DNA Fragment Detection. The goal of this research is to enhance the sensitivity of trace DNA fragment detection. Toward this end we have devised and recently demonstrated the detection of dsDNA fragments by using single-molecule fluorescence burst counting. A confocal detection system was used to observe fluorescence bursts from single molecules of dsDNA multiply labeled with the intercalation dye TO6. Flowing solutions of M13 DNA were first used to show that the number of bursts was linear with concentration, that the average burst duration was consistent with the expected transit time, and that the number of detected bursts was consistent with the concentration. The optimized single molecule apparatus and analysis method was then used to detect CE separations of M13, pBR 322 and pRL 277 DNA. Separations are easily detected when only 50-100 molecules of DNA per band pass through the detection region, and the current detection size limit is ≥1000 bp/fragment. This new detection technology should lead to the routine analysis of DNA with an on-column sensitivity better than 100 molecules/band. Applications to cancer, bacterial, viral and trace expression detection are envisioned.

The U. C. Berkeley High Sensitivity DNA Analysis Project (directed by R. A. Mathies and A. N. Glazer) was supported by the U. S. Department of Energy under contract DE-FG-91ER61125.

1. Mathies, R. A., Scherer, J. R., Quesada, M. A., Rye, H. S. and Glazer, A. N. Laser-excited Confocal Fluorescence Gel Scanner, *Rev. Sci. Instruments* **65**, 807-812 (1994).
2. Zhu, H., Clark, S. M., Benson, S. C., Rye, H. S., Glazer, A. N. and Mathies, R. A. High-Sensitivity Capillary Electrophoresis of Double-Stranded DNA Fragments using Monomeric and Dimeric Fluorescent Intercalation Dyes, *Analytical Chemistry* **66**, 1941-1948 (1994).
3. Woolley, A. T. and Mathies, R. A. Ultra-High-Speed DNA Fragment Separations Using Microfabricated Capillary Array Electrophoresis Chips, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 11348-11352 (1994).
4 Wang, Y., Ju, J., Carpenter, B., Atherton, J. M., Sensabaugh, G. F. and Mathies, R. A. High-Speed, High-Throughput THO1 Allelic Sizing Using Energy Transfer Fluorescent Primers and Capillary Array Electrophoresis, *Analytical Chemistry* **67**, 1197-1203 (1995).
5. Ju, J., Ruan, C., Fuller, C. W., Glazer, A. N. and Mathies, R. A. Fluorescence Energy Transfer Dye-Labeled Primers for DNA Sequencing and Analysis, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 4347-4351 (1995).
6. Woolley, A. T. and Mathies, R. A. Ultra-High-Speed DNA Sequencing Using Capillary Array Electrophoresis Chips, Proceedings of the International Society for Optical Engineering-SPIE, Volume **2386**, 36-44 (1995).
7. Ju, J., Kheterpal, I., Scherer, J. R., Ruan, C., Fuller, C. W., Glazer, A. N. and Mathies, R. A. Design and Synthesis of Fluorescence Energy Transfer Dye-Labeled Primers and their Application for DNA Sequencing and Analysis, *Analytical Biochemistry* **231**, 131-140 (1995).
8. Woolley, A. T. and Mathies, R. A. Ultra-High-Speed DNA Sequencing Using Capillary Electrophoresis Chips, *Analytical Chemistry* **67**, 3676-3680 (1995).
9. Haab, B. B. and Mathies, R. A. Single Molecule Fluorescence Burst Detection of DNA Fragments Separated by Capillary Electrophoresis, *Analytical Chemistry* **67**, 3253-3260 (1995).

DETECTION OF TIN-LABELED DNA ON HYBRIDIZATION CHIPS*

H.F. Arlinghaus 1, Margaret N. Kwoka 1, K. Bruce Jacobson 2, and K. L. Beattie 3, [1]Atom Sciences, Inc. and [2]Health Sciences Research Division, Oak Ridge National Laboratory, Oak Ridge, TN and [3]Houston Advanced Research Center, Houston, TX

DNA hybridization chips were constructed by attaching 17-mer sequences from M13 DNA and T7 DNA to separate sites on a platinum film on a quartz surface. The platinum was applied in islands of ~5 mm diameter or it was spread uniformly over the surface. The DNA was attached to platinum by a newly devised chemistry. M13(-20) is a 17-mer, that is complimentary to a portion of the M13 DNA, and it was labeled with ^{118}Sn atoms[1]. Hybridization of the DNA on the chip with this 17-mer, and washing, occurred at 5° C.

The position of the tin label was identified with Sputter Initiated Resonance Ionization Spectroscopy (SIRIS), a technique that analyzes a fraction of a monolayer of the surface of any sample. In the case of platinum islands, the ^{118}Sn was found to be located on all of the complementary sites and none was present on the non-complimentary sites or the intermediary space. However, the tin was not well localized on the even film of platinum, indicating that a thin layer of tin had deposited on this surface. In contrast to ^{32}P or fluorescent labels, SIRIS detects the tin present in only a fraction of a monolayer of the surface. The amount of material that spreads out non-specifically probably would not be detectable by autoradiography or fluorescence but is apparent with SIRIS.

Successful use of SIRIS and tin-labeled DNA offers an alternative to fluorescence for analyzing hybridization chips. The signal-to-noise ratio of tin-labeled DNA was over 100 to 1 whereas the signal-to-noise of fluorescent labels is often less than 4 to 1, unless confocal microscopy is employed. SIRIS has the potential to analyze several hundred separate hybridization spots per second and thus offers a rapid means to obtain data from the hybridization chip. Furthermore, several isotopes of tin or other elements can be used on different DNA probes simultaneously to provide a rapid multiplex analysis of several DNAs at once.

*Research sponsored by the Laboratory Directed Research and Development Program of the Oak Ridge National Laboratory, managed for the U. S. Department of Energy by Lockheed Martin Energy Systems, Inc., under contract No DE-AC05-84OR21400 and two SBIR Phase I's to Atom Sciences, Inc. from NIH: 1-R43-CA66525-01 and 1-R43-MH52938-01.

1. K.B. Jacobson and H.F. Arlinghaus, Anal. Chem. **64,** 315A-328A (1992)

MASS SPECTROMETRY IN THE HUMAN GENOME PROJECT

Peter Williams, Chau-Wen Chou, David Dogruel, Jennifer Krone, Kathy Lewis, Randall Nelson, Department of Chemistry and Biochemistry, Arizona State University, Tempe, AZ 85287-1604

There are three potential roles for mass spectrometry relevant to the Human Genome Project:

a) The most obvious role is that on which all groups have been focussing -- development of an alternative, faster sequence ladder readout method to speed up large-scale sequencing. Progress here has been difficult and slow because the mass spectrometry requirements exceed the current capabilities of mass spectrometry even for proteins, and DNA presents significantly more difficulty than proteins. We have shown previously that pulsed laser ablation of DNA from frozen aqueous films has the potential to yield sequence-quality mass spectra, but that ionization in this approach is erratic and uncontrollable. We are focussing on developing ionization methods using ion (or electron) attachment to vapor-phase DNA (ablated from ice films) in an electric field-free environment; results of this approach will be reported.

b) Mass spectrometry may not ultimately compete favorably in speed with large-scale multiplexing of conventional or near-term technologies such as capillary electrophoresis. However, as the Genome project nears completion there will be an increasing need for rapid small-scale DNA analysis, where the multiplex advantage will not be so great and mass spectrometry could play a more significant role there. With this in mind we are looking at ways to speed up the overall mass spectrometric analysis, e.g. simple rapid cleanup of sequence mixtures, and at generation of short sequence ladders by exopeptidase digestion.

c) Given the genome data base(s) at the completion of the project, with rapid search capability, a need will arise for comparably rapid generation of search input data to identify often very small quantities of proteins isolated from biochemical investigations. With this in mind we have developed extremely rapid enzyme digestion techniques optimized for mass spectrometric readout, using endopeptidases covalently coupled directly to the mass spectrometer probe tip. The elimination of autolysis and transfer losses allows rapid (few minute) endopeptidase digestion and mass analysis of as little as 1 picomole of protein, leading to an umbiguous database identification. An alternative search procedure uses partial amino-acid sequence information. With the added use of exopeptidases to generate a peptide ladder sequence in the mass spectrum of the endopeptidase digest, on the order of a dozen residues of internal sequence can be generated in a total analysis time of 20 minutes or less, again using only picomoles of sample.

A SELF-ASSEMBLED MATRIX MONOLAYER FOR UV-MALDI MASS SPECTROMETRY

Stéphane Mouradian, Christine M. Nelson and Lloyd M. Smith
University of Wisconsin, Department of Chemistry, Madison, WI 53706

MALDI Mass Spectrometry has brought significant advances in the analysis of large biopolymers. However, the desorption/ionization mechanism remains unclear and the particular influence of the crystal formation process has yet to be understood. In this work, we replace the usual matrix crystals by a monolayer of a higly absorbing matrix-like chemical which is covalently linked to a gold surface. In this experiment, analytes such as proteins or oligonucleotides are directly deposited on the covalently modified probe tips. The samples are irradiated at 355nm and the molecular ions are detected using a Time of Flight Mass Spectrometer (TOFMS). Several types of monolayers have been investigated and tested for their ability to produce molecular ions in the positive and negative ion modes. dT10 oligonucleotide and proteins as large as cytochrome C were successfully analyzed using this procedure. As a control, no molecular ion was detected when the analyte was deposited onto a clean bare gold surface. By further characterizing the monolayers used in this type of experiment, we expect to gain a better understanding of the mechanism of production of molecular ions in MALDI.

This work was supported by Department of Energy Human Genome grant DE-FG02-91ER61130

ANALYZING SEQUENCING REACTIONS FROM BACTERIOPHAGE M13 BY MALDI MASS SPECTROMETRY

Stéphane Mouradian, David R. Rank and Lloyd M. Smith
Department of Chemistry, University of Wisconsin, Madison WI 53706

The current demand for improved DNA sequencing methodologies posed by the Human Genome Project has spurred the investigation of alternatives to gel electrophoresis. Matrix assisted laser desorption ionization (MALDI) mass spectrometry has great potential for the rapid analysis of DNA fragments. Mock Sanger sequencing mixtures have already been successfully analyzed by MALDI by pooling synthesized oligos corresponding to the M13 bacteriophage sequence (1). More recently, the analysis of enzymatically generated Sanger sequencing fragments have been performed (2,3). However, in both these studies, a synthetic template of 45 and 50 bases was used instead of the DNA used in conventional sequencing. We have adressed this issue by performing MALDI analysis of sequencing mixtures using bacteriophage M13 as a template. Current results allow sequence determination of at least 30 bases with a 17mer primer. A typical cycle sequencing protocol (sequiTherm^{TM}) has been modified to yield about 0.4 pmol of each extension product. Different desalting and purification procedures have been investigated and it was found that salt accumulation could be efficiently reduced by removal of template salt in a post reaction step.
Work in progress (see abstract "Factors Influencing DNA Stability in Matrix Assisted Laser Desorption/Ionization (MALDI) Mass Spectrometry" by Christine Nelson *and al.*) to stabilize DNA by chemical modification employed in conjunction with methods described here, should enable significant extension of the length of readable sequence.

This work was Supported by Department of Energy Human Genome grant DE-FG0291ER61130.
(1) M.C. Fitgerald, L. Zhu, L. M. smith (1993) Rapid. Commun. Mass Spectrom. 7, 895-897.
(2) T. A. Shaler, Y. Tan, J. N. Wickham, K. J. Wu and C. H. Becker (1995) Rapid Comm. Mass Spectrom. 9, 942-947.
(3) M. T. Roskey, I. P. Sminov, P. Juhasz, M. Vestal, E. J. Talkach, S. A. Martin, L. A. Haff (1995) *7th International Genome Sequencing and Analysis Conference, September 16-20, 1995, Hilton Head S.C. Genome Science & technology (1995) 1, p46..*

Factors Influencing DNA Stability In Matrix-Assisted Laser Desorption/Ionization (MALDI) Mass Spectrometry

Christine M. Nelson, Lin Zhu, Wei Tang and Lloyd M. Smith
Department of Chemistry, University of Wisconsin, Madison, WI 53705

The development of MALDI and its demonstrated performance with large proteins, up to 300,000 daltons in size, has generated substantial interest in utilizing this technique as a replacement for the gel electrophoretic separation step in the Sanger sequencing protocol. If successful, the main advantages of this method over traditional gel-based methods are that polyacrylamide gels are no longer necessary and that the time for separation, detection and data acquisition is substantially reduced.

In developing this strategy, early studies of oligonucleotides have elucidated issues inherent to the analysis of nucleic acids using MALDI. These studies have shown that fragmentation is an important issue and is responsible for some of the current limitations associated with the technique.

Results in our laboratory and in others demonstrate that fragmentation is dependent on both oligonucleotide sequence and matrix composition.[1] Proposed fragmentation pathways consist of nucleobase protonation inducing base loss followed by backbone cleavage at the 3' C-O bond on the corresponding deoxyribose. A thorough study recently published from our laboratory elucidates this fragmentation mechanism and explains reasons for observed differences in base composition.[2] Studies in our laboratory investigating positive ion formation of oligonucleotides also support this mechanism. Currently, we are investigating the relationship between fragmentation propensity and matrix properties. Proton affinities of five common MALDI matrices have been measured and compared to fragmentation probabilities.

In addition to studying the fragmentation mechanism, we are using our current understanding to explore nucleotide modifications that may enhance the stability of DNA in the MALDI technique. Modifications on both the nucleobase and deoxyribose ring are being investigated. Experimental and theoretical approaches are being employed for this study.

By deepening our understanding of the fundamental chemistry of DNA fragmentation, we hope to be able to develop the MALDI technique into a pwerful robust and versatile methodology for nucleic acid analysis.

1. Parr, G.R.; Fitzgerald, M.C.; Smith, L.M.; *Rapid Commun. Mass Spectrom.* **7**, 63 (1993).
2. Zhu, L.; Parr, G.R.; Fitzgerald, M.C.; Nelson, C.M.; Smith, L.M.; *J. Amer. Chem. Soc.* **117**, 6048 (1995).

MASS SPECTROMETER OLIGONUCLEOTIDE ANALYZER AND ITS APPLICATIONS

C. H. Winston Chen, Nelli I. Taranenko, Y. F. Zhu, and S. L. Allman, Health Sciences Research Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6378

Laser desorption mass spectrometry has been considered as a potential method for fast DNA analysis and sequencing. In order to achieve this goal, the following four conditions need to be satisfied. They are (1) the capability of detecting DNA longer than 300 base pairs, (2) detection sensitivity for each size of DNA reaches one femtomole or less, (3) resolution good enough to resolve one base difference in size between two DNA segments, and (4) automation to achieve fast sample preparation and data analysis.

During the past two years, we have demonstrated the capability of detecting oligonucleotides of 500 base pairs in size by the discovery of new matrices and improving the mass spectrometer. Detection sensitivity has reached 100 femtomole in the samples. The resolution ($M/\Delta M$) for oligonucleotides less than 40 mer has reached 800. Sequencing a small segment of DNA by mass spectrometry is under extensive investigation.

In addition to the effort on sequencing, some effort was also put to demonstrating the use of mass spectrometry for disease diagnosis. Detection of ΔF508 with CTT deletion for cystic fibrosis has been successfully demonstrated for 30 patients samples. Point mutation related to cystic fibrosis and lung disease was also detected by mass spectrometry.

Basic understanding of matrix-assisted laser desorption/ionization was also under extensive study. Several new matrices for DNA were discovered. New approaches using ion trap configuration and pulsed ion extraction were also pursued for resolution improvement.

Details will be presented in the meeting.

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

Mass Spectrometric Molecular Weight Measurement of PCR-Amplified DNA*

Gregory B.Hurst[a], Mitchel J. Doktycz[b], R.P. Woychik[c], Michelle V.Buchanan[a], [a]Chemical and Analytical Sciences Division, [b]Health Sciences Research Division, and [c]Biology Division, Oak Ridge National Laboratory, Oak Ridge TN 37831.

The polymerase chain reaction (PCR) produces many copies of a targeted DNA sequence, with information encoded in the size of the targeted sequence. The currently-accepted method for determining PCR product sizes is by gel electrophoresis, a lengthy and labor-intensive process that is prone to inaccuracies. As a potentially faster and more accurate alternative to gel electrophoresis, we are developing mass spectrometric methods based on matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF-MS) for measuring PCR product molecular weights. Because the sensitivity of MALDI-TOF-MS decreases as the size of the analyte increases, it is desirable to design PCR assays that yield products containing 200 or fewer base pairs.

To illustrate the applicability of mass spectrometry for PCR product detection, an assay was devised for amplifying a 75-bp product corresponding to bases 1626 to 1701 of the cystic fibrosis transmembrane conductance regulator gene. The most common mutation of this gene is the ΔF508, which is a three-base deletion that produces a 72-bp PCR product. Following PCR and a simple cleanup procedure, the product was detected by MALDI-TOF-MS. As a second example, PCR products were prepared from the genomic DNA of bacteria of the genus *Legionella* using primers that are available in a commercial kit. The resulting products--a 108-mer common to all members of the genus, and a 168-mer specific to *L. pneumophila*, the organism that causes Legionnaire's disease--were detected by MALDI-TOF-MS. Current challenges for MALDI-TOF analysis of PCR products include improved methods for removing salts and other reagents, and improvements in resolution and upper mass limit.

Sensitivity is also an active area of research for MALDI-TOF. The majority of a typical sample (1 μL of a 1 μM solution of DNA or protein) can be recovered following the MALDI measurement, suggesting that the analysis can be performed with much smaller sample sizes. We are currently developing sample preparation techniques using nanoliter and smaller volumes.

*Research supported by the Office of Health and Environmental Research, U.S. Department of Energy, and the Oak Ridge National Laboratory Director's Research and Development Program, under contract DE-AC05-84OR21400 with Lockheed Martin Energy Systems, Inc.

RAPID, ACCURATE DETECTION OF DNA SEQUENCE VARIANTS VIA ELECTROSPRAY M.S.*

Chris E. Hopkins, Mark Leppert, Jim McCloskey, Pamela Crain, and Raymond Gesteland, Department of Human Genetics, University of Utah, Salt Lake City, Ut, 84112

Electrospray Mass Spectrometry (EMS) enables very accurate molecular weight determinations, to less than 0.01% error. In recent years EMS has enabled the discovery of novel and rare modifications of ribonucleic acids, and has enabled mass validation of t-RNAs up to 120 bases long. EMS analysis is rapid, typically within ten minutes per assay, and reliable, due accurate mass determination.

The EMS technique has great potential for detecting single base differences in genomic PCR amplification products. A primer pair is designed to flank a polymorphic locus. The primers, separated by one to four bases, will generate upon amplification one product from each chromosome. In a heterozygote the two PCR products will differ by at least the identity of one base pair. The mass analysis of the denatured mixture will give the molecular weight of the four strands and thus define the identity of the polymorphism.

To demonstrate this capability, PCR EMS has been applied to identifying a known C to T polymorphism in the alpha 4 subunit of the nicotinic acytylcholine receptor gene (CHα4), a candidate gene for several forms of human epilepsy. A 53 base pair region encompassing this polymorphism was amplified from a pair of primers, 24 and 25 bases in length. The 4 bases between the primers contain the C to T polymorphic site. Mass analysis of this site on a heterozygotic individual generates mass data on each of the strands amplified from the paired chromosomes. The accuracy is within 0.01% mass error compared to theoretical molecular weight, thus allowing unambigous identification of the polymorphism.

Synthetic templates have been generated to model all possible point polymorphisms at this locus. Mass data generated from these models demonstrates that all are readily detectable. This technique will enable sequence variant determination to made in hours instead of days and it should be applicable to any PCR amplified product up to a theoretical and practical limit of 70 base pairs for point polymorphisms and up to 150 for base deletion polymorphisms. It is expected that this technique will be a valuable tool for DNA diagnostic analysis.

*Supported by a grant from the Office of Health and Environmental Research of the U.S. Department of Energy under contract DE-FG94ER61817

Analysis of Oligonucleotide Mixtures by Electrospray Ionization-Mass Spectrometry

David C. Muddiman, Xueheng Cheng, Harold R. Udseth and Richard D. Smith*

Chemical Sciences Department, Pacific Northwest Laboratory, Richland, WA 99352

Our aim is to develop electrospray ionization mass spectrometry (ESI-MS) methods for high speed DNA sequencing of oligonucleotide mixtures, that can be integrated into an effective overall sequencing strategy. ESI produces intact molecular ions from DNA fragments of different size and sequence with high efficiency [1]. Our aim is to determine mass spectrometric conditions that are compatible with biological sample preparation and that avoid problems due to dissociation, aggregation, or adduction during ionization of the DNA fragments. Oligonucleotide ions are typically produced from ESI with a broad distribution of net charge states for each molecular species (i.e., $(M\text{-}nH)^{n-}$, where n is a series of integers), and thus leading to difficulties in analysis of complex mixtures [1]. To make identification of each component in a sequencing mixture possible, the charge states of molecular ions can be reduced by manipulating the ESI process and/or by using gas-phase reactions. The charge-state reduction methods being examined include: (1) reactions with organic acids and bases (in the solution to be electrosprayed and the ESI-MS interface or the gas phase); (2) the labeling of the oligonucleotides with a designed functional group for production of molecular ions of very low charge states; and (3) the shielding of potential charge sites on the oligonucleotide *phosphate/phosphodiester* groups with polyamines (and the subsequent gas-phase removal of the neutral amines). In initial studies two methods for charge state reduction of gas phase oligonucleotide negative ions have been tested: (1) the addition of acids and bases to the oligonucleotide solution and (2) the formation of diamine adducts followed by dissociation in the interface region [2]. In the first method, the efficiency of charge state reduction depends on the pK_a, the concentration and the nature of the acids. Acetic and formic acids were found to be better reagents than HCl, CF_3CO_2H and H_3PO_4; however, suppression of the analyte solution was observed. If the infused solution contained a high percentage of organic solvent, signal suppression was obviated. In addition, the addition of organic bases reduced cation adduction **and** unexpectedly reduced charge-states. The second method has the advantage that the stability of oligonucleotides is not affected but requires the optimization of the interface dissociation conditions and the amounts of diamine added to the oligonucleotide solution which may not be analytically reproducible. Both methods show promise for charge state reduction and results have been demonstrated for several small oligonucleotides (*i.e.*, $d(pT)_{12}$, d(AGCT), $d(pT)_{18}$, $d(pC)_{12}$, $d(pA)_6$, and 8-mers of A, C and T). [2,3]. Substantial reduction in the spectral density was observed for a three, four and six-component mixture of oligonucleotides sprayed from a solution containing a charge state reducing agent. Our aim is to provide a basis for the development of an overall approach to high speed sequencing to provide a basis for the subsequent step of prototyping a cost effective high-throughput instrument for broad application.

[1] "New Developments in Biochemical Mass Spectrometry: Electrospray Ionization", R. D. Smith, J. A. Loo, C. G. Edmonds, C. J. Barinaga, and H. R. Udseth, **Anal. Chem.**, 62, 882-889 (1990).

[2] "Charge State Reduction of Oligonucleotide Negative Ions from Electrospray Ionization", X. Cheng, D. C. Gale, H. R. Udseth, and R. D. Smith, **Anal. Chem.**, 67, 586-593 (1995)

[3] "Charge-State Reduction with Improved Signal Intensity of Oligonucleotides in Electrospray Ionization Mass Spectrometry" D.C. Muddiman, X. Cheng, R.D. Smith, **J. Am. Soc. Mass Spectrom.**, submitted.

This work was supported through the U.S. Department of Energy. Pacific Northwest Laboratory is operated by Battelle Memorial Institute through Contract No. DE-AC06-76RLO 1830.

High Speed Sequencing of Single DNA Molecules in the Gas Phase by FTICR-MS

Richard D. Smith*, David C. Muddiman, Xueheng Cheng, S. A. Hofstadler, J. E. Bruce
Chemical Sciences Department and Environmental and Molecular Sciences Laboratory, Pacific Northwest Laboratory, Richland, WA 99352

This project is aimed at the development of a totally new concept for high speed DNA sequencing based upon the analysis of single (i.e., individual) large DNA fragments using electrospray ionization (ESI) combined with Fourier transform ion cyclotron resonance (FTICR) mass spectrometry. In our approach, large single-stranded DNA segments extending to as much as 25 kilobases (and possibly much larger), is first transferred to the gas phase using ESI. The multiply-charged molecular ions are trapped in the cell of an FTICR mass spectrometer, where one or more *single ion(s)* can be selected for analysis in which its mass-to-charge ratio (*m/z*) is measured both rapidly and non-destructively. Single ion detection is achievable due to the high charge state of the electrosprayed ions and the unique sensitivity of the FTICR mass spectrometer developed at Pacific Northwest Laboratory.

Our efforts under the first several years of this project have demonstrated the capability for the formation, extended trapping, isolation, and monitoring of sequential reactions of highly charged DNA molecular ions with molecular weights well into the megadalton range [1-5]. We have shown that large multiply-charged individual ions of both single and *double-stranded* DNA anions can also be efficiently trapped in the FTICR cell, and their mass-to-charge ratios measured with very high accuracy. Thus, it is feasible to quickly determine the mass of each lost unit as the DNA is subjected to rapid reactive degradation steps. Our aim is to now develop methods based upon the use of ion-molecule or photochemical processes that can promote a stepwise reactive degradation of gas-phase DNA anions. Successful development of one of these approaches could greatly reduce the cost and enhance the speed of DNA sequencing, potentially allowing for sequencing *DNA segments of more* than 25 kilobase in length, on a time-scale of minutes with negligible error rates, with the added potential for conducting many such measurements in parallel. *The techniques* being developed promise to lead to a host of new methods for DNA characterization, potentially extending to the size of much larger DNA restriction fragments (>500 kilobases).

[1] "Trapping, Detection and Reaction of Very Large Single Molecular Ions by Mass Spectrometry," R. D. Smith, X. Cheng, J. E. Bruce, S.A. Hofstadler and G.A. Anderson, **Nature**, 369, 137-139 (1994).

[2] "Charge State Shifting of Individual Multiply-Charged Ions of Bovine Albumin Dimer and Molecular Weight Determination Using an Individual-Ion Approach," X. Cheng, R. Bakhtiar, S. Van Orden, and R. D. Smith, **Anal. Chem.**, 66, 2084-2087 (1994).

[3] "Trapping, Detection, and Mass Measurement of Individual Ions in a Fourier Transform Ion Cyclotron Resonance Mass Spectrometer,: J.E. Bruce, X. Cheng, R. Bakhtiar, Q. Wu, S.A. Hofstadler, G.A. Anderson, and R.D. Smith, **J. Amer. Chem. Soc.**, 116, 7839-7847 (1994).

[4] "Direct Charge Number and Molecular Weight Determination of Large Individual Ions by Electrospray Ionization-Fourier Transform Ion Cyclotron Resonance Mass Spectrometry", R. Chen, Q. Wu, D.W. Mitchell, S.A. Hofstadler, A.L. Rockwood, and R. D. Smith, **Anal. Chem.**, 66, 3964-3969 (1994).

[5] "Trapping, Detection and Mass Determination of Coliphage T4 DNA (1.1×10^8 Da) Ions by Electrospray Ionization Fourier Transform Ion Cyclotron Reasonance Mass Spectrometry" R. Chen, X. Cheng, D.W. Mitchell, S.A. Hofstadler, A.L. Rockwood, Q. Wu, M.G. Sherman and R.D. Smith, **Anal. Chem.**, 67, 1159-1163 (1995).

This work was supported through the U.S. Department of Energy. Pacific Northwest Laboratory is operated by Battelle Memorial Institute through Contract No. DE-AC06-76RLO 1830.

ADVANCED DETECTORS FOR MASS SPECTROMETRY

W.H. Benner and J.M. Jaklevic

Ernest Orlando Lawrence Berkeley National Laboratory, University of California, Engineering Science Department, Human Genome Group, 1 Cyclotron Road, Berkeley, CA 94720, Phone: (510) 486-7194, Fax: (510) 486-5857, whbenner@lbl.gov

Mass spectrometry is an instrumental method capable of producing rapid analyses with high mass accuracy. When applied to genome research, it is an attractive alternative to gel electrophoresis. At present, routine DNA analysis by mass spectrometry is seriously constrained to small DNA fragments. Contrasted to other mass spectrometry facilities in which the development of ladder sequencing is emphasized, we are exploring the application of mass spectrometry to procedures that identify short sequences. This approach helps the molecular biologists associated with LBL's Human Genome Center to identify redundant sequences and vector contamination in clones rapidly, thereby improving sequencing efficiency. Biological presequencing procedures designed to use mass spectrometry as a detection scheme will be presented as ways to perform oligonucleotide ligation assays, end-of-fragment sequencing and the sizing of deletion series created in P1, BAC and PAC vectors.

We are also working to improve the operation of matrix-assisted-laser-desorption-ionization (MALDI) and electrospray mass spectrometers by developing new ion detectors. One of the limitations for applying mass spectrometry to DNA analysis \relates to the poor efficiency with which conventional electron multipliers detect large ions, a problem most apparent in MALDI-TOF-MS. To solve this problem, we are developing alternative detection schemes which rely on heat pulse detection. The kinetic energy of impacting ions is converted into heat when ions strike a detector and we are attempting to measure indirectly such heat pulses. We are developing two detectors based on impact detection. We have generated particle impact signals in metal-oxide-silicon structures and a piezoelectric film.

Electrospray ion sources generate ions of megadalton DNA with minimal fragmentation, but the mass spectrometric analyses of these large ions usually leads only to a mass-to-charge distribution. If ion charge were known, actual mass data could be determined. To address this problem, we are developing a detector that will simultaneously measure the charge and velocity of individual ions. Mass spectra of megadalton DNA will be presented showing the feasibility for sizing fragments in the 2 to 40 kb region.

This work was supported by the Director, Office of Energy Research, Office of Health and Environmental Research, Human Genome Program, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

NEW TECHNOLOGIES FOR GENOME SEQUENCING AND ANALYSIS*

George M. Church, *Richard Baldarelli, James Chou, Poguang Wang, Helena Graner, Linxiao Xu, Pete Estep, Dereth Phillips, Saeed Tavazoie, and Keith Robison,* Department of Genetics, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02115, Church@Rascal.Med.Harvard.Edu

DNA sequencing for genome projects, diagnostics (microbial, agricultural, human, immunogenetics and oncogenetics) and industrial QC will push well beyond the 3 Gbp and 0.1% error rates originally targeted for one human genome. To increase accuracy and efficiency, of data acquisition two new technologies are in development: 1) An automated multiplex sequencer is being built to use 400 mass-spectral tags for primers per electrophoretic lane (in collaboration with Bruker Instruments, Genome Therapeutics, and Dr. Roger Giese's laboratory at Northeastern University). 2) Conductance measurements of DNA moving through single ion channels, such as lambda phage clone DNA injected into LamB receptor pores. To increase the accuracy of microbial genome sequence checking and database annotation we are developing methods for 3) computational genome comparisons in putative non-coding regions, 4) multiplex homologous recombination knockout analyses, 5) whole genome *in vivo* footprinting, and 6) locating point mutations using mismatch recognition proteins.

*Supported by a grant from the U.S. Department of Energy DE-FG02-87ER60565

Preparation of Oligonucleotide Arrays for Hybridization Studies and Arrayed Primer Extension (APEX) for Mutation Detection and Sequencing

Michael C. Pirrung, Lara Fallon, J.-C. Bradley, William P. Hawe, and Shin Han

P. M. Gross Chemical Laboratory, Department of Chemistry, Box 90346, Duke University, Durham, NC 27708.

Glenn McGall

Affymetrix, 3380 Central Expressway, Santa Clara, CA 95051.

This project is aimed at developing reliable, high-quality chemical synthetic methods to prepare high-density arrays containing thousands of short DNA sequences with either the 5' or 3' end free. Such arrays can be used for hybridization or primer extension reactions. Arrays consisting of complete sets of DNA of a given length can be prepared by the novel technique of light-directed synthesis, and photoremovable groups are key to this method. We have developed a superior new photoremovable group for light-directed DNA synthesis that gives a byproduct that is chemically inert and readily measured by optical and fluorescence methods, permitting the yield in each of the photochemical deprotection steps to be verified. The 3',5'-dimethoxybenzoin (DMB) protecting group is used for the phosphotriester method of DNA synthesis. The 3',5'-dimethoxybenzoincarbonate (DMB-carbonate) photoremovable protecting group is used for the phosphoramidite method of DNA synthesis. Complete sets of the four protected (allyl, allyloxycarbonyl) nucleoside monomers in which the DMB carbonate is on the 5'-oxygen and the phosphoramidite is on the 3'-oxygen (and *vice versa*) have been prepared. The former have been used in light-directed synthesis of arrays to both prove the chemistry and test hybridization conditions. The DMB-carbonate method has also been applied to phosphoramidite synthesis of DNA tethered to the solid phase through its 5' end. Work with a collaborating group (J. Shumaker, Baylor) has shown that arrays of DNA with this design can be effectively used for single nucleotide terminating primer extensions (using DNA polymerase and ddNTPs) templated by analyte DNA, permitting comparison sequencing. Finally, this method has been applied to mRNA templates using reverse transcriptase, permitting the analysis of gene expression.

Publications

Michael C. Pirrung, Lara Fallon, Steven W. Shuey, and David C. Lever, "Inverse Phosphotriester DNA Synthesis Using Photochemically-removable Dimethoxybenzoinphosphate Protecting Groups," J. Org. Chem., **60**, 0000 (1995).

Michael C. Pirrung and Jean-Claude Bradley, "Comparison of Methods for Photochemical Phosphoramidite-based DNA Synthesis," J. Org. Chem., **60**, 6270 (1995).

Michael C. Pirrung and Jean-Claude Bradley, "Dimethoxybenzoin Carbonates: Photochemically-removable Alcohol Protecting Groups Suitable for Phosphoramidite-based DNA Synthesis," J. Org. Chem., **60**, 1116 (1995).

Michael C. Pirrung and Steve W. Shuey, "Photoremovable Protecting Groups for Phosphorylation of Chiral Alcohols. Asymmetric Synthesis of Phosphotriesters of (-)-3',5'-Dimethoxybenzoin," J. Org. Chem., **59**, 3890 (1994).

PIEZOELECTRIC SENSOR ARRAYS FOR THE DETECTION AND QUANTIFICATION OF GENETIC POLYMORPHISM

J. C. Andle, D. J. McAllister, J. T. Weaver, C. P. H. Vary* and J. F. Vetelino**

BIODE, Inc., 4 Rockland Ct., Brewer, ME 04412-1254
*Maine Medical Center Research Institute, 125 John Roberts Rd., Suite 8
So. Portland, ME 04106

**Dept. of Electrical and Computer Eng. And Laboratory for Surface Science and Tech., 5764 Sawyer Research Center, University of Maine, Orono, ME 04469-5764

While many DNA detection applications ideally require "single copy" sensitivity, the ability to reliably amplify DNA sequences via polymerase chain reaction (PCR) techniques allows low copy numbers of target DNA to be amplified and detected using available instrumentation. Typically, the target is defined by two flanking sequences (primers) and any sequence flanked by these two primers is amplified. In practice, gel electrophoresis is employed to verify the molecular weight of the amplificant, providing further verification of sequence. However, gel electrophoresis does not allow the researcher to detect polymorphisms of a gene, in which point mutations, deletions or additions do not substantially alter the molecular weight of the DNA sequence.

One may employ a third DNA sequence - internal to the target sequence and containing the region of the suspected mutation - to perform an affinity-based recognition of the target. Typically, point mutations have dramatic effects on the affinity of DNA probes for the corresponding target. By employing an array of probe sequences - each corresponding to an anticipated mutation of the gene and immobilizing each probe to a separate detector - it is possible to differentially detect the various mutations.

One class of detectors which hold promise for such arrays are the piezoelectric biosensors [1]. Piezoelectric biosensors offer the ability to selectively detect small quantities of specific DNA - currently on the order of a few nanograms per milliliter of solution - in the presence of overwhelming concentrations of nonspecific DNA - as much as 1000-fold excess. Typically, the detection process is complete within a few minutes. These sensors directly detect the added mass of the analyte as it hybridizes with complementary DNA probes or triplex probes which are covalently bound to the surface. Mass sensitivities of a few hundred picograms per mm^2 are reported.

Past work has been directed at detecting chemically denatured DNA using complementary probes. Current effort is directed at the use of peptide nucleic acid (PNA) probes for triplex capture. Phase I of this project will investigate a four element (three probes plus reference) detector array. A DNA sample will be evaluated for the level of cross-reactivity to the "wrong" detectors relative to the "proper" and "reference" detectors.

Work supported in part by DOE under grants DE-FG02-92ER-81350 and DE-FG02-95ER-81933.

[1] "Acoustic Wave-Based Biosensors", **Invited paper**, J. C. Andle and J. F. Vetelino, Sensors and Actuators A 44, pp. 167-176 (1994).

BIOLOGICAL MICROCHIPS: DEVELOPMENT AND APPLICATIONS*

A. Mirzabekov, G. Yershov, V. Barsky, E. Timofeev, D. Guschin, V. Shick, S. Dubiley, D. Pobedimskaya, and D. Prudnikov, Argonne National Laboratory (U.S.A.) and Engelhardt Institute of Molecular Biology (Moscow) — Joint Human Genome Program.

The development of biological microchips will allow us to collect and analyze significant amounts of biological information in comparatively simple experiments. Oligonucleotide microchips have been manufactured and applied for analysis of DNA sequences. The microchips (40×40×20 μm or larger size) consist of gel elements with immobilized oligonucleotides fixed on a glass surface. A robot was constructed for large-scale microchip production, in which it applies activated oligonucleotide solutions (1–50 nl) to gel elements. A simple method was also developed for manual microchip manufacturing. The hybridization of fluorescently labeled DNA with a microchip is monitored in real time at several wavelengths, and with a temperature gradient by means of a specially devised fluorescent microscope coupled with a CCD camera. Alternatively, the hybridized unlabeled DNA can be stained directly on the microchip with a fluorescent dye. Contiguous stacking hybridization has been developed to raise the sequencing efficiency of, for example, an 8-mer microchip to that of a 13-mer microchip. The equipment and the method developed were applied for analysis of DNA sequences and as diagnostics for genetic diseases. The development of SHOM for mapping, as well as partial and complete sequencing of DNA will be presented in posters.

Our technology allows us to manufacture microchips bearing different immobilized compounds of biological interest, such as various oligonucleotides, peptides, DNA, RNA, proteins, antibodies, and low molecular weight ligands. Potential applications for such biological microchips will be discussed.

*Work supported in part by the U.S. Department of Energy, Office of Health and Environmental Research, under Contract No. W-31-109-ENG-38 and Russian Human Genome Program.

ROBOT FOR PRODUCING OLIGONUCLEOTIDE MICROCHIPS

G.M. Yershov, A.I. Belgovsky, L.D. Drobyshev, V.N. Sushkov, N.V. Mologina, D. Guschin, J. Steele, A. Gemmel, A. Zaslavsky, D. Naylor,[1] and A.D. Mirzabekov, Argonne National Laboratory (U.S.A.) and Engelhardt Institute of Molecular Biology (Moscow)—Joint Human Genome Program. [1]University of Illinois at Chicago (U.S.A.)

The manufacture of matrices with gel-immobilized oligonucleotides (sequencing microchips) to be used in SHOM (sequencing by hybridization with oligonucleotide microchips) includes three separate stages. The first stage is shaping the desired topology of the gel micromatrix by mechanical scribing, laser ablation, or photopolymerization. The second stage is to load microvolumes of oligonucleotide solutions onto the matrix of the gel "cells", and the third is to immobilize the oligonucleotides within the cells. The production rate of such matrices is limited mainly by the step of loading microvolumes.

We have designed two production robots for rapid microdispensing of the oligonucleotides which have to be immobilized onto the cells of microchips. The first robot was established in Moscow and successfully tested. It filled five cells per minute using one pin and 80 cells per minute using 16 pins. The other robot, developed at Argonne National Laboratory under the joint project, has higher productivity and accuracy. It can fill approximately 160 cells per minute using 16 pins and 640 cells per minute using 64 pins. Productivity can be increased by producing many similar microchips in parallel. The development of a fully automated line is well under way. Major features of this work will be highlighted in a poster session at the conference.

Using three-dimensional gels as carriers of oligonucleotides, microchip manufacturing technology and hardware have been developed. This approach can also be successful (with or without chemical modification of various gels) for the detection of specific chemical interactions of water-soluble bioorganic compounds. This subject will also be discussed in the poster session.

*Work supported by the U.S. Department of Energy, Office of Health and Environmental Research, under contract No. W-31-109-ENG-38; by the Affymetrix, Inc., U.S.A.; and the Russian Human Genome Program.

CHEMICAL WAYS TO IMPROVE DNA SEQUENCING BY HYBRIDIZATION WITH OLIGONUCLEOTIDE MICROCHIPS (SHOM)*

Edward N. Timofeev, Andrew G. Kunitsyn, Svetlana V. Kochetkova, Andrew D. Mirzabekov, and Vladimir L. Florentiev, Argonne National Laboratory (U.S.A.) and Engelhardt Institute of Molecular Biology (Moscow) — Joint Human Genome Program.

Sequencing DNA by using the hybridization with oligonucleotide microchips (SHOM) method is one of the prospective methods for DNA sequencing and mutation detection.[1,2] To improve this method, we have carried out the following investigations: (i) development of more effective procedures for oligonucleotide immobilization (ii) development of an approach to increase the selectivity of hybridization and (iii) study of the physico-chemical basis of DNA hybridization on the microchip.

To increase the stability of oligonucleotide binding to the gel, we have elaborated a number of alternative methods for activation of the gel. These techniques are based on copolymerization of acrylamide with acrylic acid derivatives containing either hydrazide or amino groups. Immobilization of oligonucleotides on copolymers was carried out by the three following methods: (**a**) interaction of hydrazide copolymer with an dialdehyde derivative of the oligonucleotide; (**b**) treatment of amino copolymer with an dialdehyde derivative of the oligonucleotide in the presence of reductant (pyridine-borane complex), which results in the formation of a very stable amino bond; and (**c**) reaction of amino copolymer with an oligonucleotide derivative containing an activated ester (*N*-hydroxysuccinimide ester) coupled through a linker to oligomers. In the latter case, the binding of oligomers is stronger than in method **a** but no as strong as in method **b**.

To increase the selectivity of hybridization, we have studied the hybridization properties of oligonucleotides in which the universal base analog 5-nitroindole was added to the end of the sense sequence. Hybridization experiments showed that discrimination between perfect duplexes and duplexes with terminal mismatches was increased.

Earlier we proposed a theoretical model for the dissociation of the duplex formed by DNA with immobilized oligonucleotides.[3] To apply this model practically for the calculation of relatively stability of duplexes, we have performed an experimental study of the thermodynamic parameters of such duplex formation. The results obtained enable us to prepare the "normalized" matrix (the matrix with equalized stability of duplexes having GC-contents).

*Work supported in part by the U.S. Department of Energy, Office of Health and Environmental Research, under Contract No. W-31-109-ENG-38 and Russian Human Genome Program.

[1]Khrapko, K.R., Lysov, Yu.P., Khorlin, A.A., Shick, V.V., Florentiev, V.L., Mirzabekov, A.D. *FEBS Lett.*, **256**, 118–122 (1989).
[2]Khrapko, K.R., Lysov, Yu.P., Khorlin, A.A., Ivanov, I.B., Yershov, G.M., Vasilenko, S.K., Florentiev, V.L., and Mirzabekov, A.D. *J. DNA Sequencing and Mapping*, **1**, 375–388 (1991).
[3]Livshits, M.A., Florentiev, V.L., and Mirzabekov, A.D. *J. Biomol. Struct. Dynam.* **11**, 783–795 (1994).

SHOM: SIGNIFICANT IMPROVEMENT IN THE ACCURACY OF DNA HYBRIDIZATION TESTS BY THE STATISTICS OF MISMATCHED DUPLEXES*

M.A. Livshits, and A.D. Mirzabekov, Argonne National Laboratory (U.S.A.) and Engelhardt Institute of Molecular Biology (Moscow) — Joint Human Genome Program.

The goal of a DNA hybridization test with a set of octanucleotides is to reveal the eight-letter words present in the DNA text. The straightforward identification of perfect duplexes is subject to a number of inevitable random errors. To diminish the errors, one can use, instead of statistics of repeated measurements, rather large statistics of data provided by a single hybridization experiment with the complete octanucleotide matrix. To this end, one should take into account also mismatched duplexes. The presence of a given word in the DNA sequence is confirmed not only by the perfect duplex formed with the octanucleotide strictly complementary to the word, but also by the obligatory formation of single-mismatch duplexes with all 24 related single-substitution octanucleotides. If, on the other hand, the octanucleotide fixed in a given hybridization cell forms with DNA a single-mismatch duplex, then in most of related single-substitution cells (21 or 24), double-mismatch duplexes with the same DNA site will be formed. Because of this, the "family test" evaluating the signal from a given hybridization cell together with the signals from 24 related single-substitution cells deals in fact with the difference in stability of single- and double-mismatch duplexes rather than with identification of a perfect duplex. Importantly in such a test, a much higher discrimination power can be achieved due to the fact that the ratio of standard deviation to mean value is considerably lower (by 4.6 times) for the "family" signal than for the individual cell signals.

*Work supported in part by the U.S. Department of Energy, Office of Health and Environmental Research, under Contract No. W-31-109-ENG-38 and Russian Human Genome Program.

EFFICIENCY OF SEQUENCING BY HYBRIDIZATION ON OLIGONUCLEOTIDE MATRIX SUPPLEMENTED BY MEASUREMENT OF THE DISTANCE BETWEEN DNA SEGMENTS*

Yuri P. Lysov, Fedor N. Gnuchev, Andrei A. Mironov, Alexey A. Chernyi, and Andrei D. Mirzabekov, Argonne National Laboratory (U.S.A.) and Engelhardt Institute of Molecular Biology (Moscow) — Joint Human Genome Program.

DNA sequencing by hybridization on oligonucleotide microchips (SHOM) allows the determination of a spectrum of overlapping oligonucleotides constituting a DNA fragment. The oligomers that hybridize to form perfect duplexes with an array of immobilized oligonucleotides and, as a result, reconstitution of the nucleotide sequence of the fragment is possible. In longer DNA fragments, unambiguous reconstitution of a DNA sequence is often impeded by the presence of repetitive regions and simple sequence repeats. Here it is demonstrated that SHOM, supplemented by measurement of the distance between certain sites within the analyzed DNA (for example, restriction sites or priming sites for PCR), enables sequencing of much longer DNA fragments containing repeats of varying complexity. Results of computer simulations on sequences from the EMBL database in the context of a model experiment including contiguous stacking hybridization are presented. These results show expected efficiency of the method as function of the sequence origin and are intended to be used as guidelines for real sequencing experiment planning. Sequences up to 10,000 bp long are reconstructed efficiently.

*Work supported in part by the U.S. Department of Energy, Office of Health and Environmental Research, under Contract No. W-31-109-ENG-38 and Russian Human Genome Program.

DETECTION OF BETA-THALASSEMIA MUTATIONS BY HYBRIDIZATION WITH OLIGONUCLEOTIDE MICROCHIPS

V. Shick, S. Dubiley, N. Kalganova, D. Pobedimskaya, D. Prudnikov, A. Perov, D. Guschin, S. Belikov and A. Mirzabekov, Argonne National Laboratory (U.S.A.) and Engelhardt Institute of Molecular Biology (Moscow) — Joint Human Genome Program.

Oligonucleotide microchips have been tested as diagnostics of genetic hybridization diseases. Several methods have been evaluated for the preparation of hybridization probes. Double-stranded DNA shows much lower efficiency of hybridization with the microchips than does single-stranded DNA. The efficiency is increased significantly by fragmentation of the dsDNA into short pieces. Different methods of preparation of ssDNA have been tested, such as asymmetric PCR, isolation of PCR biotin-ssDNA on Dynabeads, and immobilization of denatured PCR dsDNA on a resin followed by DNA synthesis with a DNA primer. Long ssDNA fragments may form hairpin structures and diffuse slowly into the gel to interact with the gel-immobilized oligonucleotides; they are poorly hybridized with the microchips. DNA fragmentation significantly increased the hybridization efficiency. Enzymatic and chemical random fragmentation of DNA have been compared. Chemical and enzymatic methods of introducing fluorescent labels into DNA fragments for monitoring of the hybridization on the microchips by fluorescent microscopy have been developed. The use of PCR with T7 RNA promoter containing primers and the following transcription with T7 RNA polymerase have also been tested for preparing RNA for microchip hybridization. The effect of different conditions of hybridization in a low-volume hybridization chamber have been tested. A reliable identification of heterozygous and homozygous beta-thalassemia mutations has been demonstrated by hybridization with the microchips of the probes prepared by these methods.

*Work supported in part by the U.S. Department of Energy, Office of Health and Environmental Research, under Contract No. W-31-109-ENG-38 and Russian Human Genome Program.

This page intentionally left blank.

Mapping

Creating Physical Maps from Hybridization Data: Why it's easy, Why it's hard*

David S. Greenberg, Cynthia A. Phillips, Sandia National Laboratories, Mail Stop 1110, P.O. Box 5800, Albuquerque, NM 87185-1110 and David Wilson, Massachusetts Institute of Technology

There have been many studies which investigate algorithms for reconstructing physical maps from hybridization data between clone libraries and sets of genomic probes[6, 8, 4, 3, 1]. All of the various formulations of the problem of finding a complete physical map have been shown to be NP-complete and thus a variety of heuristic and/or approximate solutions have been proposed. Interestingly, the analyses of these approaches (whether by theoretical bounds, evaluation against known maps, or evaluation against simulated data) have given mixed results. Some studies imply that excellent maps are possible while others show that good maps are unlikely.

In [5] an attempt was made to formalize the model of the problem so that everyone could agree on the conditions involved. In this work we use the formalization to examine the expected quality of the maps. We start by looking at the simple case in which the hybridization data is error free – that is, each clone represents a single contiguous section of genome and all hybridizations are faithfully recorded.

Surprisingly to us, there is still are large amount of inherent ambiguity in maps created from such data. As a measure of ambiguity we asked what percentage of unique probes (STSs for example) which are actually adjacent on the genome could have been non-adjacent on a genome for which the hybridization data remained unchanged. We dubbed these adjacencies, *weak*, since no algorithm can be sure they are adjacent based only on the hybridization data.

We found that three factors effect the amount of weakness in the result, the coverage of the clones, the length of the clones, and the statistical manner in which the clones and probes were chosen. The effects of coverage have been previously reported in studies such as [7, 2]. Our new results are that clone length has a non-linear effect on the results. Once a critical length (between 3 and 5 depending on the manner in which clones are closen) is reached the amount of weakness reaches an asymptote. On the other hand, very small expected clone lengths result in very high weakness. We evaluated several models of clone and probe generation. The method of clone generation varied between constant size, size bounded in a range, and Poisson distributed sizes. While expected length was the most important factor, allowing the sizes to be Poisson distributed increased the expected weakness somewhat. On the other hand, the method of probe generation was varied from uniform gaps to Poisson distributed gaps. The use of Poisson distributed gaps led to significantly greater weakness.

We conclude that the reason for the varied results of analyses of algorithms for physical mapping is that the method in which clones and probes are generated varies greatly. We suggest that more careful analysis in which the ambiguity of the result is compared against our theoretical result will yield more reliable judgements about algorithms.

* This work was supported by the U.S. Department of Energy and was performed at Sandia National Laboratories for the U.S. Department of Energy under contract DE-AC04-94AL85000.

References

[1] F. Alizadeh, R. Karp, L. Newberg, and D. K. Weiser. Physical mapping of chromosomes: a combinatorial problem in molecular biology. In *Proceedings of the 4th Annual ACM-SIAM SODA*, pages 371–381, 1993.

[2] R. Arratia, E. S. Lander, S. Tavare, and M. S. Waterman. Genomic mapping by anchoring random clones: A mathematical analysis. *Genomics*, 11:806–827, 1991.

[3] A. Cuticchia, J. Arnold, and W. Timberlake. The use of simulated annealing in chromosome reconstruction experiments based on binary scoring. *Genetics*, 132:591–601, 1992.

[4] A. Cuticchia, J. Arnold, and W. Timberlake. ODS: ordering DNA sequences – a physical mapping algorithm based on simulated annealing. *CABIOS*, 9(2):215–219, 1993.

[5] D. Greenberg and S. Istrail. Physical mapping by STS hybridization: Algorithmic strategies and the challenge of software evaluation. *JCB*, 2(2):219–274, 1995.

[6] A. Grigoriev, R. Mott, and H. Lehrach. An algorithm to detect chimeric clones and random noise in genomic mapping. *Genomics*, 22:282–486, 1994.

[7] E. S. Lander and M. S. Waterman. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, 2(3):231–329, 1988.

[8] R. Mott, etal. Algorithms and software tools for ordering clone libraries: application to the mapping of the genome *schizosaccharomyces pombe*. *Nucleic Acid Research*, 21(8):1965–1974, 1993.

PROGRESS IN HIGH DENSITY GRIDDED ARRAY HYBRIDIZATION

A. Copeland (copeland2@llnl.gov),*D. Masquelier, R. Langlois, J. Kimbrough, L. Mascio, B. Pesavento, R. Mariella, E. Branscomb*. Human Genome Center, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory,Livermore, CA 94550

Researchers at LLNL and throughout the world continue to make use of high density gridded arrays of transformed bacteria or DNA to support sequencing and gene finding efforts. We are using a commercial robotic positioning system to supplement our Hewlett-Packard ORCA robot, which we have used previously for producing gridded arrays. The new system is capable of producing usable colony arrays at densities approaching 100,000 spots on an 8-cm x 12-cm substrate. We have also developed a 384-pin tool using pins from the semiconductor industry which is more accurate than our previous 384-pin tool. The system is currently being used in a production mode to make arrays at a 6 x6 x384 density using this 384-pin tool. These lower-density arrays are suitable for analysis using radio-labeled probes and storage phosphor data capture. Our immediate target is 10 x10 x384 arrays of BAC clones analyzed with fluorescently-labeled probes. The new spotting robot has a work envelope of 0.25m x 1m x 2m and allows us to produce up to 164 8-cm x 12-cm filters at a time.

Work is also in progress on a computer aided image analysis system which is interfaced to our genome data base to provide both direct image recovery and automated recording of results. Toward the goal of using fluorescent-based detection methods in very high density array contexts, we are investigating new visible and infra-red labeling strategies and the use of alternative spotting substrates with lower fluorescence backgrounds and improved target presentation. We also describe improvements made to the colony growth, lysis and fixing protocols, including fabrication of a fixture that allows more efficient batch processing of colony filters.

(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.)

COMPARATIVE HYBRIDIZATION TO ARRAYS OF DNA CLONES

D. Pinkel[1,2], Donna Albertson[1,3], Y. Zhai[2], R. Segraves[2], D. Sudar[1], K. Ligtenberg[1], and J. Gray[1,2].
[1]Lawrence Berkeley National Laboratory, [2]University of California San Francisco, [3]MRC Laboratory of Molecular Biology

Many genomic and genetic studies face the challenge of searching for differences in gene dosage or expression among cell populations. For example malignancies may involve the loss or gain of DNA sequences or abnormal expression levels of particular genes, reproductive abnormalities frequently result from loss or gain of chromosome segments, differential gene expression underlies development, and genetics frequently requires mapping duplicated or deleted regions of mutant genomes. Several years ago comparative genomic hybridization (CGH), which permits surveying entire genomes for regions of variant gene dosage was developed. In current practice two genomes, for example a test and a normal reference genome, are labeled with different fluorochromes and simultaneously hybridized to normal metaphase chromosomes. Regions of the test genome with elevated or reduced copy number are indicated by corresponding variations in the ratio of the hybridization signals along the target chromosomes. This technique has proven very powerful, but it has substantial limitations that stem from the use of metaphase chromosomes as the hybridization target. The most fundamental of these is that the genomic resolution is at best several Mb due to the complex packing of DNA in metaphase chromosomes. This determines the precision with which copy number variations can be mapped and the sensitivity with which changes affecting a small region can be detected.

To address these problems we are developing the ability to make arrays of cloned DNA that are suitable for comparative fluorescence hybridizations with probes of total mammalian genomic complexity. Currently each element of the array contains approximately 100 pg of target DNA in a 50-100 μm diameter spot on a glass microscope slide. Probes are labeled with fluorescein or Texas red. Test hybridizations in a model system consisting of lambda DNA targets hybridized with red and green labeled lambda DNA probes in various ratios indicate that the ratio of the fluorescence intensities is accurately proportional to the ratio of the two probe concentrations over a range of more than 10^3 in relative concentration. Signals can be detected with probe concentrations down to 2 pg/μl, which is equivalent to the concentration of a 50 kb length of DNA in a CGH hybridization involving mammalian genomic DNA. The amplification of chromosome 20q in breast cancer cell line BT474 was detected using a 4 element array containing total human genomic DNA, a P1 clone from chromosome 20q13, a P1 clone from chromosome 18, and lambda DNA. The fluorescence ratio on the 20q13 target was 2-4 times greater than that on the chromosome 18 and total genomic DNA targets. No hybridization was detectable on the lambda target. These measurements indicate the potential for using arrayed targets for quantitative comparisons of relative concentrations of multiple nucleic acid sequences in high complexity mixtures.

Supported by a grant from the director, Office of Energy Research and Development, Office of Health and Environmental Research, Department of Energy, under contract DE-AC-03-76SF00098; Vysis Inc. by subcontract from National Institute of Science and Technology, Department of Commerce; and NIH grant CA45919.

TOWARDS A FULL-LENGTH cDNA LIBRARY: A PROGRESS REPORT

Marcelo Bento Soares, Kala Mayur, Maria de Fatima Bonaldo and Susan Baumes
Department of Psychiatry, Columbia University and The New York State Psychiatric Institute, New York, NY 10032

We have further optimized our original procedure for construction of normalized directionally cloned cDNA libraries [1] and we have successfully applied it to generate human cDNA libraries from fetal liver-spleen, full-term and 8-9 week placentae, adult breast, brain, retina, pineal gland, ovary tumor, melanocytes and multiple sclerosis plaques. Several additional libraries are currently in preparation. All libraries have been contributed to the IMAGE consortium, and they are being widely used for sequencing and mapping.

Several milestones have been accomplished towards our final objective of generating full-length normalized cDNA libraries. First, we have adapted our normalization protocol to take advantage of the fact that it is now possible to produce single-stranded circles *in vitro* by sequentially digesting supercoiled plasmids with Gene II protein and Exonuclease III (Life Technologies). This is significant because it circumvents the biases introduced by differential growth of clones containing small and large cDNA inserts when single-strands are produced *in vivo* upon superinfection with a helper phage. Second, a number of parameters have been optimized regarding synthesis of first and second strand cDNA. Third, we have optimized conditions for size selection of RNA and we have generated several size fractions of a pool of placentae and brain mRNAs. We are currently starting to construct cDNA libraries from individual RNA size fractions. The fact that the size range of the starting template RNA population is known makes it possible that we perform a strict size selection of the synthesized double-stranded cDNAs and thereby greatly enrich for full-length molecules prior to cloning.

Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy under contract DE-FG02-91ER61233

[1] Soares, M.B., Bonaldo, M.F., Su, L., Lawton, L. & Efstratiadis, A. Construction and characterization of a normalized cDNA library (1994). Proc. Natl. Acad. Sci. USA **91**(20), 9228-9232.

The I.M.A.G.E. Consortium

Greg Lennon[1], *M. Bento Soares*[2]. [1]Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94550 and [2]Dept. of Psychiatry, Columbia University, New York, NY 10032. [1]Corresponding author.

Founded with Drs. Auffray (CNRS) and Polymeropoulos (NCHGR) after discussions at a previous DOE Contractor's- Grantee meeting, the I.M.A.G.E. Consortium is a collaborative effort to systematically identify the majority of genes through the use of arrayed cDNA libraries. Sequence, map, and expression data derived from the clones in these libraries is placed in public databases, and the clones themselves are available royalty-free. Over 200,000 clones have been arrayed at LLNL from 22 human cDNA libraries, and over 200,000 associated EST sequences have been deposited in dbEST primarily through the efforts of the WashU-Merck collaboration. Five organizations worldwide now distribute I.M.A.G.E. clones and associated reagents (such as high-density hybridization filters). In conjunction with efforts such as the Merck Gene Index to determine the number of distinct genes represented (currently over 30,000), I.M.A.G.E. clones are substrates for genome wide transcriptional mapping and full-insert sequencing to complement high-throughput genomic sequencing. We are also conducting subtraction experiments to enrich for clones representing remaining undiscovered genes. Large-scale characterization of arrayed cDNAs from other species is also likely to enhance our knowledge not only of gene number but, more importantly, of both gene diversity and function. For further information, contact the Consortium by e-mail (info@image.llnl.gov) or through the WWW (http://www-bio.llnl.gov/bbrp/image/image.html).

(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.)

Hybrid Selection of cDNAs from 1 Megabase of Human Chromosome 19

Wufang Fan[1], *Greg Lennon*. Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94550. [1]Corresponding author.

The goal of this effort is to use cDNA hybrid selection techniques to isolate, sequence, and map coding regions located on human chromosome 19. Our initial experiments used flow-sorted chromosome 19 DNA and arrayed cDNA libraries, and have resulted in cDNA sublibraries enriched approximately 10 fold for sequences from chromosome 19. To increase the enrichment of the target genes, we are now selecting cDNAs hybridizing to cosmid contigs from the physical map of chromosome 19 (see other posters for physical map details).

Two contigs spanning in total 1 megabase have been studied. A contig of 400 kb located within 19p12-13.1, and containing the gene defective in pseudoachondroplasia and multiple epiphyseal dysplasia, was the first on which our protocols were optimized. Pooling the results from selection experiments with three different cDNA libraries, 55% of the selected cDNA fragments map back specifically to the starting contig. Even though the cosmids lack ribosomal sequences, a major contaminant of the remaining 45% of selected clones are sequences homologous to rRNA. After sorting the cDNAs based on both hybridization and sequencing, 17 distinct genes have been characterized. We have also compared these 17 genes with the results of an exon-trapping experiment conducted in parallel using the same cosmids. A second contig spanning 600 kb of chromosome 19q13.1 has recently been used as the basis for hybrid selection experiments. By choosing cDNA material with low amounts of rRNA contamination, we have been able to increase the percentage of selected fragments mapping back to the starting cosmids to 65-75%, even after only one round of selection. These results indicate that the rate-limiting steps in isolating genes from a region are now increasingly the ability to identify different cDNA fragments derived from the same gene, and the isolation of corresponding full-length cDNAs.

(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.)

Towards a globally integrated, sequence-ready BAC map of the human genome

Ung-Jin Kim, Hiroaki Shizuya, and Melvin I. Simon
Division of Biology, 147-75, California Institute of Technology, Pasadena, CA 91125

BACs and Fosmids are stable, non-chimeric, highly representative cloning systems. The BACs maintain large fragment genomic inserts (100-300 kb)[1] and their DNA is easily prepared for most types of experiments including DNA sequencing[2]. We have been improving BAC cloning techniques and constructed > 10X human and mouse BAC libaries. As BACs are proving to be the most efficient reagents for genomic sequencing, we intend to increase the depth of the library up to 30X genomic equivalence to be able to construct optimal contig maps from which one could select minimally overlapping BAC sets for genomic sequencing.
The possibility of using BACs as a generalized tool to build a global physical map was explored in our on-going chromosome 22 mapping project. Approximately 700 mapped markers including cDNAs, ESTs, STSs, cosmids, Fosmids, and other landmarks were used to screen the first 4X library. The density of the landmarks in this approach was approximately 1 per every 50-60 kb stretch of chromosome 22q. Many of these markers have been ordered on the YAC-based framework map[3], allowing rapid and precise localization of BAC contigs along the long arm of chromosome 22. Over 80% of the chromosome has been covered by BACs that have been identified and mapped to corresponding loci by markers. We currently have more than 1,000 chromosome 22-specific BACs, or on the average 3X coverage of chromosome 22q, which are now being characterized by restriction fingerprint analysis and the extent of overlaps between the clones in the contigs determined. Closure of gaps is being sought by screening deeperBAC library with markers and BAC end probes.

Currently large numbers of human genes that have been discovered and exist in the form of sequence-tagged cDNAs or ESTs are being assigned to genomic subregions via YACs and radiation hybrids. Because the landmarks from the YAC framework map have allowed rapid assembly of BAC maps on the chromosome 22q arm, it is feasible to employ the ESTs from the radiation hybrid/YAC frameworks as landmarks and rapidly assemble BACs to generate genome-wide BAC contig maps. Approximately 30,000 such landmarks will correspond to a density of 1 landmark in less than 100 kb of euchromatin. We are planning to utilize initially 30,000 mapped ESTs or cDNAs to construct BAC contigs on the entire genome. The resulting BAC-EST maps, even before its completion, will provide high resolution EST (or gene) maps, and more importantly, entry points for gene finding and large scale genomic sequencing.

*Supported by a Department of Energy grant # FG0389ER60891.
1. Shizuya, H., Birren, B., Kim, U.-J., Mancino, V., Slepak, T., Tachiiri, Y., and Simon, M.I. (1992) Proc.Natl. Acad. Sci. USA 89, 8794-8797.
2. Kim,U.-J., Birren,B.W., Yu-Ling Sheng, Tatiana Slepak, Valeria Mancino, Cecilie Boysen, Hyung -Lyun Kang, Melvin I. Simon, and Hiroaki Shizuya, submitted.
3. Collins,J.E. et al. (1995) Nature, in press.

BACs, PACs and the Structure of the Human Genome

[1]J. R. Korenberg, [1]X-N. Chen, [1]S. Mitchell, [1]Z. Sun, [1]E. Vataru, [2]U-J. Kim, [4]P. de Jong, [2]M. Simon, [3]T. J. Hudson, [3]B. Birren, [3]E. Lander, [3]J. Silva, [3]X. Wu. [1]Cedars-Sinai Research Institute, L.A., CA; [2]Caltech, Pasadena, CA; [3]Whitehead Institute/MIT, Boston, MA; [4]Roswell Park Cancer Institute, Buffalo, NY.

Not all that glitters is single copy sequence. In order to study genome organization and to provide an integrated, genomic framework for gene isolation, sequencing and mapping, we have established a Mapped BAC/PAC Resource. The goal is to represent unequivocally, 0.8-1.2X of the human genome in a stable framework resource, integrated at 1-5,000 loci with the RH, genetic and STS maps.

Current Resource: Human

The current Mapped BAC/PAC Resource now defines 4,300 sites, and represents about 18% of the human genome. We have assigned 4,000 of 17,000 BAC/PAC clones, including 3750 BACs and 250 PACs, to regions of 2-6 Mb by using fluorescence in situ hybridization, and have integrated 91 BACs with the genetic, YAC/STS, and radiation hybrid (RH) maps by using PCR of 1,000 markers to screen the 17,000 BACs. More than 250 sites are non-tandem repetitive sequence sites; 264 BAC/PACs recognize alpha satellite sites, of which 143 were selected with a consensus alpha oligonucleotide, 102 are specific to a single chromosome and the totality of all alpha-BACs now recognize all chromosomes except 10 and the Y. Finally, BACs selected by a TTAGGG consensus oligonucleotide recognize 18 telomeres, 5 of which are specific to a single chromosome.

Information on the Resource, is available on the WWW site http://www.csmc.edu/genetics/korenberg/korenberg.html that includes request forms and agreements. To facilitate distribution, screening, and aneuploidy applications, 2,902 BACs were rearrayed to reflect the true chromosomal organization, from chromosome 1p through 22q.

Mouse

Using high resolution techniques, 100 BACs have been mapped to single bands in the mouse genome.

Human Disease and Genome Organization

Analysis of the 227 BACs on chromosome 7 suggests a novel genomic structure involving clustered low-copy repetitive sequences whose arrangement likely predisposes to the deletions responsible for Williams syndrome. Similar clustering of other subsets of BACs suggests that this structure may be a model for the existence of additional subsets of low copy interspersed repeated sequences that account not only for deletions responsible for human disease syndromes but also for a subset of somatic deletions and rearrangements responsible for cancers.

The Mapped BAC/PAC Resource now provides rapid approaches to genome organization and a rapidly integrated and flexible framework for mapping and sequencing the human genome.

LARGE HUMAN AND MOUSE PAC LIBRARIES FOR PHYSICAL MAPPING AND GENOME SEQUENCING, AND MORE VERSATILE CLONING VECTORS*

Joe Catanese[1], *Baohui Zhao*[1], *Eirik Frengen*[1], *Chenyan Wu*[1], *Xiaoping Guan*[1], *Chira Chen*[1], *Eugenia Pietrzak*[1], *Panayotis A. Ioannou*[2], *Julie Korenberg*[3], *Joel Jessee*[4] *and Pieter J. de Jong*[1], [1]Department of Human Genetics, Roswell Park Cancer Institute, Buffalo, NY 14263, [2]The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus, [3]Cedar Sinai Medical Center, Los Angeles, CA 90048,[4]Life Technologies, Gaithersburg, MD 20898.

Recently, we have developed procedures for the cloning of large DNA fragments using a bacteriophage P1 derived vector, pCYPAC1 (Ioannou *et al.* (1994), Nature Genetics **6:** 84-89). A slightly modified vector (pCYPAC2) has now been used to create a 15-fold redundant PAC library of the human genome, arrayed in more than 1,000 384-well dishes. DNA was obtained from blood lymphocytes from a male donor. The library was prepared in four distinct sections designated as RPCI-1, RPCI-3, RPCI-4 and RPCI-5, respectively, each having 120 kbp average inserts. The RPCI-1 segment of the library (3X; 120,000 clones, including 25% non-recombinant) has been distributed to over 40 genome centers worldwide and has been used in many physical mapping studies, positional cloning efforts and in various large-scale DNA sequencing enterprises. Screening of the RPCI-1 library by numerous markers results in an average of 3 positive PACs per autosome-derived probe or STS marker. In situ hybridization results with 250 PAC clones indicate that chimerism is low or non-existing. Distribution of RPCI-3 (3X, 78,000 clones, less than 1% non-recombinants, 4% empty wells) is now underway and the further RPCI-4 and -5 segments (< 5% empty wells) will be distributed upon request. To facilitate screening of the PAC library, we have provided the RPCI-1 PAC library to several screening companies and non-commercial resource centers. In addition, we are now distributing high-density colony membranes at cost-recovery price, mainly to groups having a copy of the PAC library. The combined RPCI-1 and -3 segments (6X) can be represented on 11 colony filters of 22x22 cm, using duplicate colonies for each clone. We are currently generating a similar PAC library from the 129 mouse strain.

To facilitate the additional use of large-insert bacterial clones for functional studies, we have prepared new PAC & BAC vectors with a dominant selectable marker gene (the blasticidin gene under control of the beta-actin promoter), an EBV replicon and an "update feature". This feature utilizes the specificity of Transposon Tn7 for the Tn7att sequence (in the new PAC and BAC vectors) to transpose marker genes, other replicons and other sequences into PACs or BACs. Hence, it facilitates retrofitting existing PAC/BAC clones (made with the new vectors) with desirable sequences without affecting the inserts. The new vector(s) are being applied to generate second generation libraries for human (female donor), mouse and rat.

*Supported in part by grants from the Office of Health and Environmental Research of the U.S. Department of Energy (#DE-FG02-94ER61883) and the National Center for Human Genome Research, National Institutes of Health (#1R01RG01165).

INCREASING THE INFORMATION CONTENT OF STS-BASED GENOME MAPS: IDENTIFYING POLYMORPHISMS IN MAPPED STSs*

Pui-Yan Kwok 1, Qiang Deng 2, Hamideh Zakeri 1, Scott L. Taylor 2, and Deborah A. Nickerson 2, 1 - Division of Dermatology, Washington University School of Medicine, St. Louis, MO 63110, 2 - Department of Molecular Biotechnology, University of Washington School of Medicine, Seattle, WA 98195.

Numerous groups are engaged in the physical mapping of the human genome by constructing STS-content maps. STS mapping has been the dominant method by which large, well validated clone-based maps have been constructed. More than 30,000 STSs, with the average spacing of 100 kb, will be available when STS-content maps of the human genome are completed. It is interesting to note that although thousands of well-characterized and physically mapped STSs are available already, no attempts to screen these STSs for DNA variations have been reported to date. Given the fact that these STSs represent hundreds of kilobase-pairs (kb) of unique human DNA sequence, and the estimate that DNA sequence variations such as single base-pair substitutions are found approximately every 1 to 2 kb, hundreds of new diallelic markers can be developed from these mapped STSs with minimal additional effort.

By screening 154 of the STSs published by the Whitehead Institute/MIT Genome Center, we have identified 45 new DNA sequence polymorphisms among the 37.2 kb of unique DNA sequence contained in these STSs, or one polymorphism every 827 bp. Forty of these variations are substitution polymorphisms (1 every 930 bp scanned) while the remaining five sequence variations are unique insertion/deletion polymorphisms (1 in 7.5 kb scanned). Using a sequence-based approach to estimate allele frequencies for these variations, 27 of the substitution polymorphisms (one in every 1.4 kb of sequence scanned) were found to have heterozygosities exceeding 32%.

Use of STS markers on the genetic map were crucial to the construction of the first generation physical maps. However, our study demonstrates that with limited investment, the genetic map can be further enhanced by developing markers from STSs on the physical map. For example, the goal of current physical mapping efforts is aimed at developing STSs ordered along a chromosome at 100 kb intervals. Assuming that 1 out of every 4 STSs are >250 bp in size, or 1 such STS every 400 kb (4 X 100 kb spacing), and that a polymorphism is identified in every 5 STSs >250 bp in size, suggesting that 1 in every 20 STSs on the physical map would be polymorphic. Thus, minimal scanning efforts on longer STSs would yield one new genetic marker every 2 Mb on the rapidly emerging physical maps of human chromosomes and would produce more than 1,500 new diallelic markers suitable for high throughput genotyping of human populations for linkage disequilibrium or allelic association studies.

* Supported by a grant from the U.S. Department of Energy under contract DE-FG06-94ER-619090.

Mapping of Cellular Senescence Genes by Functional Complementation of Tumor Cells: A Gene that Restores SEnescence in Ovarian Tumor Cells Maps to 6q16-21

Raghbir S. Athwal, Arbansjit K. Sandhu, Deepthi Reddy, Neena Deoghare and G. Pal Kaur
Fels Institute for Cancer Research and Molecular Biology, Temple University School of Medicine, Philadelphia, PA 19140

We have concluded the production of mouse/human monochromosomal hybrid cell lines. The current panel is comprised of chromosomes 1-3, 5-17, 19-21, X and Y. In addition to the monochromosomal hybrid cell lines, we have also assembled a panel of normal human diploid cell lines each carrying *gpt* integrated into a different chromosome. These panels are currently used to map genes involved in cellular senescence which may be lost during the evolution of human tumors.

Human cells in culture have a limited life span, as do the cells of other species. After a number of generations normal cells display morphological changes, cessation of proliferation and senescence. In comparison many tumor cells have overcome senescence and can grow continuously in culture and *in vivo*. Thus cellular immortalization may represent a critical step in tumor progression. We have identified genes on human chromosomes 3, 6, 9, 13 and 17 which restore senescence when introduced into ovarian tumor cells. One of these genes which restores senescence in human as well in rat ovarian and breast tumor cells has been mapped to 6q16-21.

Single *gpt* tagged normal human chromosomes, present in mouse/human monochromosomal hybrid cells, were introduced into human and rat ovarian tumor cells *via* microcell fusion. Chromosome transfer clones were isolated by growth in the medium containing mycophenolic acid (25ug/ml) and xanthine (70ug/ml, MX medium) to select for *gpt*. Introduction of chromosome 3, 6, 9, 13 or 17 led to the senescence of both human and rat tumor cells while transfer of chromosomes 10 or 14 had no effect on morphology or growth potential of these cells. The reappearance of tumor type cells concordant with the loss of donor human chromosome further confirmed the presence of a cellular senescence genes on these chromosomes. Immortal revertant clones also appeared among senescent cells maintained in MX medium to retain introduced chromosome. These revertant clones seem to arise due to deletions in donor chromosomes. Detailed analysis of the chromosome 6 revertant clones using microsatellite markers to assess minimum chromosomal deletion revealed that a senescence gene may be located in the region 6q16-21. Microcell transfer of a chromosomal region 6q13-21 (*gpt* tagged) into human and rat ovarian tumor cells induced cell senescence while introduction of another chromosome 6 lacking this region had no effect on cell growth. These results indicate that a gene which impart senescence to ovarian tumor cells is present in the chromosomal region 6q13-21.

REGION-SPECIFIC LIBRARIES FOR THE HUMAN GENOME*

Fa-Ten Kao and Jingwei Yu, Eleanor Roosevelt Institute for Cancer Research, 1899 Gaylord Street, Denver, C0 80206, and Department of Biochemistry, Biophysics and Genetice, University of Colorado Health Sciences Center, Denver, CO

Since we developed the chromosome microdissection and MboI-linker adaptor microcloning techniques (1), we have constructed and characterized 11 region-specific libraries for the entire human chromosome 2: 4 libraries for the short arm and 6 libraries for the long arm, plus a library for the centromere region (2-8). Each library comprises hundreds of thousands of MboI-cleaved sequences, with a mean size of 200-250 bp. About half of the plasmid microclones contain unique sequences, and between 70 to 90% of the microclones were derived from the dissected region. In addition, we have isolated and characterized many unique sequence microclones from each library that can be readily sequenced as STSs, or in isolating other clones with large inserts for contig assembly. These libraries have been used successfully for high resolution physical and linkage mapping, and for positional cloning of disease-related genes assigned to these regions, e.g. the cloning of the gene for hereditary nonpolypsis coloretal cancer (9,10). For each library, we have established a plasmid sub-library of at least 20,000 independent microclones. These sub-libraries have been deposited to the American Type Culture Collection (ATCC) for general distribution.

Comparing to human chromosomes like 3, 4, 5, 7, 11, 12, 13, 16, 19, 21, 22 and X, chr. 2 is one of the human chromosomes that are largely under-studied, with insufficient probes and mapping details. In order to accelerate whole genome sequencing and positional cloning particularly in under-exploited chromosomes, we are constructing additional region-specific libraries for these chromosome regions, including 40 libraries in 10 chromosomes: 3 libraries for chr. 20, 3 for chr. 18, 3 for chr. 17, 3 for chr. 15, 3 for chr. 14, 4 for chr. 10, 4 for chr. 9, 4 for chr. 8, 5 for chr. 6, and 8 for chr. 1. Progress has already been made in constructing libraries for chromosomes 17, 18 and 20. A complete set of region-specific libraries for the under-studied parts of the human genome should furnish valuable resources for the genome commnuity. These libraries can be used to isolate more densely populated probes for high resolution physical mapping and contig assembly, and also for isolating region-specific cDNA clones as candidate genes in positional cloning (11,12). Moreover, the microclones with short inserts are particularly suited for large scale sequencing projects in these under-mapped regions.

*Supported by a grant from DOE (DE-FG03-94ER-61819).

[1]F.T.Kao and J.W.Yu, Proc. Natl. Acad. Sci. USA 88,1844-1848 (1991).
[2]J.Yu, et al., Genomics 14, 769-774 (1992).
[3]J.Yu, et al., Somat. Cell Mol. Genet. 20, 133-136 (1994).
[4]J.Yu, et al., Hum. Genet. 93, 557-562 (1994).
[5]J.Yu, et al., Somat. Cell Mol. Genet. 20, 353-357 (1994).
[6]F.T.Kao, et al., Cytogenet. Cell Genet. 68, 17-18 (1995).
[7]J.Yu, et al., Somat. Cell Mol. Genet. 21, 133-137 (1995).
[8]F.T.Kao, et al., submitted
[9]F.S.Leach, et al., Cell 75, 1215-1225 (1993).
[10]F.S.Leach, et al., Hum. Mol. Genet. 3, 2082 (1994).
[11]J.Yu, et al., Am. J. Hum. Genet. 51, 263-272 (1992).
[12]F.T.Kao, et al., Genomics 23, 700-703 (1994).

THE UK MEDICAL RESEARCH COUNCIL HUMAN GENOME MAPPING PROJECT AND THE ROLE OF THE RESOURCE CENTRE.

Keith Gibson,Martin Bishop,Chris Mundy.HGMP Resource Centre,Hinxton Hall,Hinxton,Cambridgeshire CB10 1RQ,UK.

An important component of the UK Human Genome Mapping Project strategy has been the provision of communal resources and services. The Resource Centre provides the community with a range of biological resources which include genomic and cDNA libraries and probe and primer banks- these are either created 'in house' or donated by individual scientists.These are then made available to 3700(as October 1995) regestered users in return for information and data on the resources used.
High throughput linkage mapping against a European Interspecific Mouse Backcross is in progress and completion is expected at the end of 1996 when some 6000 markers will be placed on the 1000 animal backcross. Data and information on the resources together with a large range of databases(including GDB)and tools are made available through the HGMP-Resource Centre computing facilities.Although the services are made available free to the UK and other European academic centres,commercial users and those outside Europe are required to pay a fee to use the resources-details are printed in G-Nome news(a news letter published by the Resource Centre).

Probe Development at the Resource for Molecular Cytogenetics*

W.-L. Kuo, C. Collins, J. Cochran, K. Greulich, D. Kowbel, J. Marstaller, K. Myambo, D. Pinkel, L. Riedell, Yu-Ping Shi, M. Wang, H.-U. Weier, P. Yue, M. Zorn, and J. Gray. LBL/UCSF Resource for Molecular Cytogenetics, Dept. Laboratory Medicine, University of California, San Francisco, CA 94143-0808 and Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720.

The LBL/UCSF Resource for Molecular Cytogenetics has been created to develop probes and associated technologies to facilitate the cytogenetic analyses. One goal of this Resource is to develop reference probes at ~5 Mb intervals spanning the human genome. These probes are mapped using fluorescence in situ hybridization (FISH), contain important genes or genetically mapped sequence tagged sites (STSs) to integrate genetic and physical maps and are sufficiently large to perform well as FISH probes in clinical material.

Human genomic libraries cloned into P1, PAC and BAC vectors are the major sources of these probes because clones from these libraries have insert sizes ranging from 75-300 kb making them well suited as molecular cytogenetic probes. Moreover, a very low frequency of chimerism and cross hybridization, less than 1%, has been observed upon mapping these clones.

To date, approximately one thousand probes have been developed or acquired by the Resource. These include probes selected for 579 specific genes or genetically mapped loci, 17 chromosome-specific centromeric probes, 31 whole chromosome paints and 202 anonymous probes. Molecular cytogenetic probes have been developed for genes known to be important. These include E-cadherin, RARA, p53, PML, HBE, TCRA, SRC, GADD45, BTK1, c-myc, sis, gli, NFKB2, NTRK1, Bcl-x, E2F-1, ERBB2, VHL, MTS-1 and genes at 20q13.2. Clones have also been selected that localize to genetic intervals implicated in Cri-du-Chat, Angelman/Prader-Willi, Langer-Giedion, Miller-Dieker, and Di George syndromes. We expect to complete development of probes to genes or genetically mapped loci spaced at ~ 5 Mb intervals of the human genome by next year.

Mouse chromosome-specific P1 clones have been developed to facilitate chromosome identification. To date, we have mapped 44 probes covering 34 loci. These probes were selected in collaboration with Dr. Eric Lander at the Whitehead Institute and Genome System to genetically mapped loci proximal to the chromosome centromere and at the telomere of each chromosome.

Information about probes developed by the Resource and their availability can now be surveyed on the Internet Web server at URL http://rmc-www.lbl.gov/.

* Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research, Department of Energy, under contract DE-AC-03-76SF00098 and Vysis, Inc.

HUMAN ARTIFICIAL EPISOMAL CHROMOSOME (HAECS) FOR CLONING, SHUTTLING AND FUNCTIONAL ASSAY OF LARGE GENETIC UNITS IN HUMAN AND RODENT CELLS

Min Wang,[1] *Panayotis A. Ioannou,*[1,2] *Michael Grosz,*[1] *Subrata Banerjee,*[1] *Evy Bashiardes,*[1,2] *Michelle Rider,*[1] *Tian-Qiang Sun*[3] *and Jean-Michel H. Vos*[1,3] - [1]Lineberger Comprehensive Cancer Center; [3]Department of Biochemistry and Biophysics; University of North Carolina, Chapel Hill, North Carolina 27599, [2]The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus. [Vos: 919 966-3036 (Phone); 919 966-3015 (Fax) and *vos@med.unc.edu* (E-mail)]

Of some 100,000 human genes, only a few thousand have been cloned, mapped or sequenced so far. Much less is known about other chromosomal regions such as those involved in DNA replication, chromatin packaging, and chromosome segregation. Construction of detailed physical maps is only the first step in localizing, identifying and determining the function of genetic units in human cells. Studying human gene function and regulation of other critical genomic regions that span hundreds of kilobase pairs of DNA requires the ability to clone an entire functional unit as a single DNA fragment and transfer it stably into human cells.

We have developed a human artificial episomal chromosome (HAEC) system based on latent replication origin of the large herpes Epstein-Barr virus (EBV) for the propagation and stable maintenance of DNA as circular minichromosomes in human cells.(1,2) Individual HAECs carried human genomic inserts ranging from 60 to 330 kb and appeared genetically stable. An HAEC library of 1500 independent clones carrying random human genomic fragments with average sizes of 150 to 200 kb was established and allowed recovery of the HAEC DNA. This autologous HAEC system with human DNA segments directly cloned in human cells provides an important tool for functional study of large mammalian DNA regions and gene therapy. (3,4)

Current efforts are focused on (a) shuttling large BAC/PAC genomic inserts in human and rodent cells and (b) packaging BAC/PAC/HAEC clones as large infectious Herpes Viruses for shuttling genomic inserts between mammalian cells and (c) constructing bacterial-based human and rodent HAEC libraries. (a) We have designed a "pop-in" vector, which can be inserted into current BAC-or PAC-based clone via site-specific integration. This "CRE-LOXP"-mediated system has been used to establish BAC/PAC up to 250 kb in size in human cells as HAECs. (b) We have obtained packaging of 160-180 kb exogenous DNA into infectious virions using the human lymphotropic Epstein-Barr virus. After delivery into human ß-lymphoblasts cells the HAEC DNA was stabley established as 160-180 kb functional autonomously replicating episomes. (5,7) (c) We have also generated a hybrid BAC/HAEC vector, which can shuttle large DNA inserts, i.e., at least up to 260 kb, between bacteria and human cells. Such a system is being used to develop large insert libraries, whose clones can be directly transferred into human or rodent cells for functional analysis. These HAEC-derived systems will provide useful molecular tools to study large genetic units in humans and rodents, and complement the functional interpretation of current sequencing efforts.

*Supported by the Office of Health and Environmental Research, Human Genome Program, Department of Energy, under Contract No. DE-FG05-91ER61135

1. Sun, T.-Q., Fenstermacher, D. & Vos, J.-M.H. Human artificial episomal chromosomes for cloning large DNA in human cells *Nature Genet* 8, 33-41 (1994).

2. Sun, T.-Q. & Vos, J.-M.H. Engineering of 100-300 kb of DNA as persisting extrachromosomal elements in human cells using the HAEC system in *Methods molec. Genet.* (ed. Adolph, K.W.) (Academic Press, San Diego, CA, 1995).

3. Vos, J.-M.H. Herpesviruses as Genetic Vectors in *Viruses in Human Gene Therapy* (ed. Vos, J.-M.H.) 109-140 (Carolina Academic Press & Chapman & Hall, Durham N.C., USA & London, UK, 1995).

4. Kelleher, Z. & Vos, J.-M. Long-Term Episomal Gene Delivery in Human Lymphoid Cells using Human and Avian Adenoviral-assisted Transfection. *Biotechniques* 17, 1110-1117 (1994).

5. Banerjee, S., Livanos, E. & Vos, J.-M.H. Therapeutic Gene Delivery in Human ß-lymphocytes with Engineered Epstein-Barr Virus. *Nature Medicine, Accepted.*

6. Sun, T.-Q., Livanos, E., & Vos, J.-M.H. Infectious HAECs for Disease Correction. *Nature Medicine,* Submitted.

7. Wang, S. & Vos, J.-M.H. An HSV/EBV based vector for High Efficient Gene Transfer to Human Cells in vitro/in vivo. *Submitted.*

SPECIFIC CLONING OF HUMAN DNA AS YACs BY TRANSFORMATION-ASSOCIATED RECOMBINATION

Vladimir Larionov, Natalya Kouprina, Joan Graves, X-N Chen[1], Julie R. Korenberg[1], and Michael A. Resnick, Laboratory of Molecular Genetics, NIEHS, Box 12233, Research Triangle Park, NC 27709, and[1] Ahmanson Department of Pediatrics, Division of Genetics, Cedars-Sinai Research Institute, UCLA, CA 90048-1869.

DNA molecules undergoing transformation into yeast are highly recombinogenic, even when diverged. We reasoned that Transformation-Associated Recombination (TAR) could be employed to clone large DNAs containing repeat sequences, thereby eliminating the need for in vitro enzymatic reactions such as restriction and ligation and reducing the amount of DNA handling. Gently isolated human DNA was transformed directly into yeast spheroplasts along with two genetically marked (*M1* and *M2*) linearized vectors that contained a human *Alu* at one end and a telomere sequence at the other end (*Alu-CEN-M1-TEL* and *Alu-M2-TEL*). Nearly all the *M1* selected transformants had YACs containing human DNA inserts that varied in size from 70 kb to more than 600 kb. Approximately half of these had also acquired the unselected *M2* marker. The mitotic segregational stability of YACs generated from one (*M1*) or two (*M1* + *M2*) vector(s) was comparable suggesting de novo generation of telomeric ends. Since no YACs were isolated when rodent DNAs or a vector lacking an Alu were used, the YACs were likely the consequence of TAR between the repeat elements on the vector(s) and the human DNA. Using the BLUR13 *Alu*-containing vector, we demonstrate that human DNA can be efficiently cloned from mouse cells that contain a single human chromosome 16. The distribution of cloned DNAs on chromosome 16 was determined by fluorescence *in situ* hybridization (FISH). We propose that TAR-cloning can provide an efficient means for generating YACs from specific chromosomes and subchromosome fragments and it may be useful for isolating families of genes and specific genes from total genome DNA.

Support was provided in part by the US Department of Energy (DE-FG03-94ER61402) to J. A. K and by an interagency grant (1-YO2-HG-60021-01) from the NIH Human Genome Center to M. A. R.

CHROMOSOME SPECIFIC COSMID AND LAMBDA LIBRARIES FOR HALF THE HUMAN GENOME

Jeffrey Garnes, Anne Bergmann, Jerry Eveleth, Benjamin Wong, Glenda Quan, Wanda Johnson, Jennifer McNinch, Jennifer Alleman, Hillary Massa, Barbara Trask, Ger van den Engh, Pieter de Jong, Jeffrey Gingrich, Richard Langlois, and Anthony V. Carrano. Human Genome Center, Biology and Biotechnology Research Program, L-452, Lawrence Livermore National Laboratory, Livermore, CA 94550

The goal of the National Laboratory Gene Library Project has been to construct chromosome-specific clone libraries for the entire complement of human chromosomes. A joint effort between Livermore and Los Alamos National Laboratories has resulted in the construction of highly redundant partial digest lambda (~15 kb) and cosmid (~40 kb) libraries for each human chromosome. Libraries constructed at Lawrence Livermore National Laboratory have been prepared for chromosomes 1, 2, 3, 7, 9, 12, 18, 19, 21, 22, X, and Y. The source of DNA for all of the libraries has been human chromosomes sorted from human/rodent hybrid somatic cell lines containing a reduced number of human chromosomes. Using High Resolution Flow Karyotype Analysis and High Speed Chromosome Sorting, millions of human chromosomes are sorted in a single day with enrichment frequencies ranging from 7X for larger chromosomes to greater than 40X for the smaller chromosomes. All of the lambda libraries have been cloned into the Charon 40 replacement vector and amplified aliquots deposited in the American Type Culture Collection, Rockville, MD. Large arrays of the 40 kb insert cosmid libraries have been generated utilizing two different cloning vectors. Lawrist5 or Lawrist16 and pFos1 are double cos-site containing vectors that facilitate the cloning of nanogram quantities of *MboI* partial digest DNA. Both vectors contain desirable features that facilitate clone propagation and chromosome mapping. We have distributed in excess of 65 copies of individual libraries to researchers throughout the scientific community and identified several resource centers in both the United States and Europe which will receive the entire collection of arrayed libraries. The construction of chromosome-specific libraries has produced an invaluable set of resources that have aided in the mapping of human chromosomes. These resources are providing sequence-ready substrates that will contribute to the ultimate map, the sequence of the human genome.

(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.)

PROGRESS TOWARDS THE CONSTRUCTION OF BAC LIBRARIES FROM FLOW SORTED HUMAN CHROMOSOMES*

Jonathan L. Longmire, Nancy C. Brown, Deborah L. Grady, Evelyn W. Campbell, Mary L. Campbell, John J. Fawcett, Phil Jewett, Robert K. Moyzis, and Larry L. Deaven, Life Sciences Division and Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545

Over the course of the National Laboratory Gene Library Project (NLGLP) we have constructed a series of DNA libraries from flow sorted human chromosomes. Small insert, complete digest libraries cloned into the EcoRI site of Charon 21A are available from the American Type Culture Collection, Rockville, MD. Partial digest libraries cloned into cosmid (sCos1) or phage (Charon 40) vectors have been constructed for chromosomes 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 20, X and Y. Purity estimates by *in situ* analysis of sorted chromosomes, flow karyotype analysis, and plaque or colony hybridization indicate that most of these libraries are 90-95% pure. Additional cosmid library constructions, 5-10X arrays of libraries into microtiter plates, and high density membrane arrays of libraries are in progress. We have also constructed a limited number of human chromosome-specific YAC libraries. In addition, we have constructed chromosome-specific M13 or pBluescript libraries for generating STS markers and for selection of chromosome-specific inserts from total genomic YAC libraries.

Because of the advantages of large insert size and stability associated with BAC cloning systems, we are currently attempting to adapt the pBelloBAC vector for use with flow sorted human chromosomes. The technical challenges involved in accomplishing this goal include developing methodologies that will allow predictable partial digestion of very small masses of DNA embedded in agarose plugs and improving BAC cloning efficiencies to allow construction of libraries from microgram quantities of chromosomal DNA. Currently, we are making modifications to pBelloBAC that include adding SacII and ClaI restriction sites into the cloning region of the vector. In addition, we have modified and significantly increased the efficiency of methods that are used to recover flow sorted chromosomes into agarose plugs prior to DNA isolation. These improvements together with new methods enabling partial digestion of chromosomal DNA samples (in progress) could allow the construction of BAC libraries from flow sorted human chromosomes.

*This work was supported by the USDOE under contract W-7405-ENG-36.

HIGH SPEED OPTICAL CHROMOSOME SORTING BASED ON LASER INDUCED PHOTOINACTIVATION OF UNWANTED CHROMOSOMAL DNA

M. C. Roslaniec, J. C. Martin, R. J. Reynolds, L. S. Cram, Los Alamos National Laboratory

We are developing a High Speed Optical Chromosome Sorter based on selective, irreversible photoinactivation of unwanted chromosomal DNA. Chromosomes will be analyzed as in conventional flow cytometry but no droplets will be generated. After analysis, unwanted chromosomes will be irradiated with a high power laser designed to impart photoinactivation. When desired chromosomes pass this photoinactivation point, the laser beam is interrupted by an optical modulator. The desired chromosomes are not photoinactivated and will subsequently be cloned. This method of chromosome sorting is an extension of the 'zapper' principle in which selectively photodamaged cells do not survive when placed in culture.[1] We expect optical chromosome selection rates of >1000 s^{-1} for analysis rates of 50,000 s^{-1}, fifty times that possible with conventional sorters.

We have successfully demonstrated photoinactivation of GM130 chromosomes in a flow cytometer. Several methods of photoinactivation are available including direct far UV damage, photosensitization via oxygen dependent and independent mechanisms or both. In our system, the DNA inactivation step is based on photoadduct formation between a psoralen derivative and chromosomal DNA. Prior to sorting, GM130 chromosomes are incubated (under dark conditions) with trimethylpsoralen. Once in the flow system, unwanted chromosome/psoralen complexes are irradiated with a UV beam from a high power argon ion laser.

Our ultimate goal is the construction of chromosome specific libraries, hence, we are using the s-Cos-1 vector to examine the effects of photodamage on cosmid cloning. Using this cosmid cloning system, the average size of packaged insert DNA is 35-45 kbp.[2] Each irradiated chromosome/psoralen complex is estimated to receive a UV exposure of ≈ 10 kJ/m^2. **With psoralen and this UV exposure, we are able to form sufficient lethal photoadducts per unit insert to reduce cloning efficiencies of GM130 chromosomal DNA to <5 % while maintaining the clonability of unirradiated DNA.**

Supported by the National Flow Cytometry Resource, NIH grant RR01315 and by the U. S. Department of Energy

(1) Keij, J. F.; Groenewegen, A. C.; Dubelaar, G. B. J.; Visser, J. W. M. *Cytometry* **1995**, *19*, 209-216. High-Speed Photodamage Cell Selection Using a Frequency-Doubled Argon Ion Laser.

(2) Evans, G. A.; Lewis, K.; Rothenberg, B. E. *Genome* **1989**, *79*, 9-20. High Efficiency Vectors for Cosmid Microcloning and Genomic Analysis.

CONSTRUCTION OF A HIGH RESOLUTION P1/PAC/BAC MAP IN THE REGION OF 5q3 FOR DIRECTED GENOMIC SEQUENCING

Jan-Fang Cheng, Steve Lowry, Duncan Scott, Yiwen Zhu and Eddy Rubin
Human Genome Center, Life Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

The LBNL Human Genome Center has focused its production genomic sequencing on the distal long arm of human chromosome 5. This region was chosen because it contains a cluster of growth factor and receptor genes and is likely to yield new and functionally related genes through long range sequence analysis.

Our mapping goal is to generate sequence-ready templates completely covering the target region. Templates would include clones isolated from human P1[1], PAC[2] and BAC[3] libraries. There are three key steps in our mapping strategy. First is the use of inter-Alu fragments generated from chromsome 5 non-chimeric YACs to isolate regionally specific clones. Second is to establish overlaps and orientations of the isolated clones. Overlaps between P1s, PACs and BACs were detected in filter hybridization using probes generated by vector-Alu PCR. Contigs were further oriented using STSs developed from known genes, ordered markers, and ends of P1s, PACs, BACs and YACs. The third step is to close gaps and verify the integrity of the cloned fragments. DNA sequences flanking gaps were used to identify additional clones for gap closure. Comparison between restriction fragments of genomic and cloned DNA allows us to sample the integrity of the cloned fragments.

Eighty-four non-chimeric YACs spanning approximately 42 Mb of the distal long arm of chromosome 5 were identified using fluorescent in situ hybridization (FISH). They formed 10 contigs which range in size from less than 2 Mb to approximately 9 Mb. These YACs were the major source of DNA for generating probes to identify P1, PAC and BAC clones in this region.

Two hundred seventy-four P1s, 53 PACs and 42 BACs were mapped to a 10 Mb region of 5q31 which contains the interleukin gene cluster and its distal 9 Mb of DNA. Of these clones, 223 P1s and all of the PACs and BACs were sized by using pulsed-field gel (PFG) electrophoresis. Chromosomal location of these clones were further confirmed by using FISH. Over 146 STSs derived from the end sequences of P1s, PACs and BACs were used to determine orientation of the clones and orientation of the contigs. The end STSs were also used in gap closure. A subset of these mapped clones were used as templates for production sequencing.

We are in the process of expanding the high resolution clone map to cover all of 5q3. Another 157 P1s, 21 PACs and several hundred BACs were isolated using probes derived from YACs located on the distal portion of 5q3. These clones have been assigned to specific regions of YACs, sized by PFG, and their location on chromosome 5 was confirmed by FISH.

Supported by a grant from the DOE under contract DE-AC03-76SF00098.

[1]Shepherd et al., *Proc. Natl. Acad. Sci. U.S.A.* 91, 2629-33 (1994).
[2]Ioannou et al., *Nature Genetics* 6, 84-9 (1994).
[3]Shizuya et al., *Proc. Natl. Acad. Sci. U.S.A.* 89, 8794-97 (1992).

A MEGA-YAC /STS PHYSICAL MAP FOR THE SHORT ARM OF HUMAN CHROMOSOME 5 *

Ellen Peterson, Donna L. Robinson, Leslie Chasteen, Robert Sutherland, Linda S. Thompson, Meryl Gersh, Joan Overhauser, Larry L. Deaven, Robert K. Moyzis, and Deborah L. Grady, Center for Human Genome Studies and Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM, and Thomas Jefferson University, Philadelphia, PA.

A total of 304 new STSs have been generated from flow sorted human chromosome 5 DNA. These STSs have been regionally ordered using breakpoint analysis to one of 51 bins on 5p and to one of 16 bins on 5q. The current density of markers (1/640 kb), in addition to the numerous PCR based genetic markers generated by other groups, is sufficient to provide nucleation points for YAC contig assembly in all regions of chromosome 5. Complete Mega-YAC contigs have been generated for the critical region (5p15.2) [Genomics 24(1):63-8, 1994] and the cat-cry region (5p15.3) identified with the cri-du-chat syndrome on the short arm of chromosome 5. This work has been extended to include YAC coverage of the entire short arm of this chromosome. The short arm constitutes approximately 50 Mb of the total 194 Mb of DNA on chromosome 5. This map has been generated by STS content mapping of YACs from the published Genethon tiling data set and by direct screening of the entire Genethon library. YACs with a high probability of containing these new STSs were collected. This probability was established by binning Genethon STSs on the same hybrid panel used on the chromosome 5 STS data. YACs localized to a specific bin by Genethon STSs were likely to contain chromosome 5 specific STSs in that same bin. By systematically testing our STSs by PCR against probable YACs and screening the library for novel YACs, a YAC contig was constructed. Currently the contig, covering >90% of 5p, consists of 531 mega-YACs and 161 integrated STSs. The completion of this map will serve as a resource for the identification of other genes on the short arm of human chromosome 5, and form the framework to construct a high resolution *E. coli* based map for eventual DNA sequencing.

*This work was supported by USDOE under contract W-7405-ENG-36.

A Highly Integrated Physical Map of Human Chromosome 16

Norman A. Doggett[1], Lynne A. Goodwin[1], Judith G. Tesmer[1], Linda J. Meincke[1], Michael R. Altherr[1], Amanda A. Ford[1], David C. Bruce[1], David C. Torney[1], Robert D. Sutherland[1], William J. Bruno[1], Emanual H. Knill[1], Grant R. Sutherland[2], David F. Callen[2], Larry L. Deaven[1], and Robert K. Moyzis[1]. [1]Los Alamos National Laboratory, Los Alamos, NM, [2]Women's & Children's Hospital, Adelaide, Australia

We have constructed an integrated map of human chromosome 16 (Doggett *et al.*, Nature 377:Suppl:335-365, 1995). The framework for constructing this map is a high resolution cytogenetic breakpoint map derived from 78 mouse/human somatic cell hybrids and 4 fragile sites which divide chromosome 16 into 90 intervals of average size 1 Mb. The physical map consists of both a low resolution YAC contig map and a high resolution cosmid contig map. The low resolution YAC contig map is comprised of 700 CEPH megaYACs, and 250 flow-sorted 16-specific miniYACs that are localized to and ordered within the breakpoint intervals with 435 STSs. This YAC map provides practically complete coverage of the euchromatic arms of the chromosome.

A high resolution "sequence ready" cosmid contig map consisting of 4000 fingerprinted cosmids assembled into contigs covering 60% of the chromosome is anchored to the YAC and cytogenetic breakpoint maps via STSs developed from cosmid contigs and by hybridizations between YACs and cosmids. The largest of these cosmid contigs spans greater than 1 million base pairs of band 16p13.3 and has been used to initiate a sample sequencing (SASE) approach to gene localization on this chromosome (see abstracts of Ricke *et al.*, Chi *et al.*).

A highly informative microsatellite-based genetic map (developed at the Adelaide Woman's and Children's Hospital) and the CEPH consortium linkage map is tightly integrated with the physical map--because nearly all of the genetic markers comprising these maps were either screened against the YAC map or localized to the cytogenetic breakpoint map.

An exon map consisting of 1000 distinct exons was recently developed from plate pools of the chromosome 16 cosmid library (see abstract of Altherr et al.). These exons are being mapped to cosmid contigs in the integrated map by hybridization of exons to cosmid grids. The integrated chromosome 16 map also includes over 600 genes, ESTs, anonymous DNA markers and microsatellite repeats--as part of an ongoing effort to incorporate all available GDB loci with the map. Supported by the US DOE under contract W-7405-ENG-36.

The Chromosome 16 Physical Map: Ordering Quality Statistics

E. Knill, Computer Research and Applications; CIC-3, MS B265, Los Alamos National Laboratory, Los Alamos, New Mexico 87545; Theoretical Biology and Biophysics Group; T10, MS K710, Los Alamos National Laboratory, Los Alamos, New Mexico 87545.

An important component of recently constructed physical maps is the ordering of STS's and other types of markers. The ordering is based on results from screening the markers against libraries of clones such as YACs and MegaYACs and localization in "bins" with tools such as somatic cell hybrid panels. The ordering of the markers usually cannot be determined uniquely from these experiments. There are two main sources of ordering ambiguities. The first is due to incomplete data. For example, markers with the same screening results cannot be distinguished. This type of ambiguity is intrinsic to the ordering problem and can be described by the use of a data structure called the PQ-tree[3]. The second source of ordering ambiguities comes from errors in the data, chimerism and hits attributable to repeated sequences. We refer to these collectively (and somewhat misleadingly) as *errors*. The purpose of this work is to give information about the second source of ambiguity.

This work contains an overview and explanation of some simple ordering quality statistics for the ordering of markers in the recently completed chromosome 16 physical map[2]. These statistics include information on changes in the number of "gaps" if adjacent sets of markers are interchanged and the number of times adjacent or nearby markers are linked by MegaYACs. The chromosome 16 physical map includes position information which we used for computing data on multiply linked pairs of markers. The position information was inferred primarily by the use of SEGMAP[4].

Because of the lack of a general agreement on or an understanding of how to best describe local ordering reliability for physical maps, we do not attempt to formally interpret the data at this time. Instead, we simply define the statistics that are presented, describe how they were obtained and tabulate them. By doing so we hope to stimulate further discussion of the issues involved and to enable future analyses of reliability information.

*This work was performed under the auspices of the U.S. Department of Energy under Contract No. W-7405-ENG-36

[1] E. Knill, The Chromosome 16 Physical Map: Ordering Quality Statistics, Los Alamos National Laboratory Report LAUR-95-2924

[2] N.A. Doggett et al. An Integrated Map of Human Chromosome 16, to appear in *Nature* (1995)

[3] K. Booth and G. S. Lueker, *Journal of Computer and System Sciences*, 13:335–379, 1976.

[4] E. D. Green and P. Green. *PCR Meth. Appl.*, 1:77–90, 1991.

DISTRIBUTION OF RESTRICTION FRAGMENTS AND REPEATS ON CHROMOSOME 16

R.D. Sutherland, R.K. Moyzis, and N.A. Doggett
Los Alamos National Laboratory, Los Alamos, NM 87545

The distribution of *Eco* RI, *Hind* III and *Eco* RI/*Hind* III double digest sites and the distribution of (GT)n and Cot1 repetitive DNA was determined for 1843 cosmids mapping to 307 locations, 75 different somatic cell hybrid breakpoint intervals, and 25 bands on chromosome 16. The average frequency of *Eco* RI, *Hind* III and *Eco* RI/*Hind* III sites per cosmid for all 1843 cosmids was 6.87, 7.34 and 9.90 respectively. The frequency of *Eco* RI, *Hind* III and *Eco* RI/*Hind* III sites in cosmids that are most likely to be located in giemsa light bands was 6.65, 7.05, and 9.66 respectively. The frequency of *Eco* RI, *Hind* III and *Eco* RI/*Hind* III sites in cosmids that are most likely to be located in giemsa dark bands was 7.18, 7.75, 10.23 respectively. The higher frequency of *Eco* RI and *Hind* III sites in giemsa dark bands is consistent with these bands having a higher A + T content than light bands since both *Eco* RI and*Hind* III recognize restriction sites in which 4 of 6 bases are A or T.

Cot1 DNA was present on an average of 5.16 *Eco* RI fragments (75%) and 4.93 *Hind* III fragments (67%) and 6.54 double digest fragments (66%) per cosmid. The frequency of Cot1 positive *Eco* RI, *Hind* III and *Eco* RI/*Hind* III fragments in giemsa light bands was 5.12, 4.83, and 6.59 respectively. The frequency of Cot1 positive *Eco* RI, *Hind* III and *Eco* RI/*Hind* III fragments in giemsa dark bands was 5.24, 5.04, and 6.41 respectively. (GT)n repeats were present on an average of 1.71 *Eco* RI fragments (25%) and 1.73 *Hind* III fragments (24%) and 1.85 double digest fragments (19%) per cosmid. The (GT)n repeats follow the same light and dark patterns as the other fragments. The averages are down ~0.10 for light bands and up ~0.10 for dark bands. These results suggest that Alu repeats (the predominate repeat in CotI DNA) are slightly more prevalent in giemsa dark bands. Supported by the US DOE (W-7405-ENG-36).

An exon based phase I expressed sequence map of chromosome 16

Michael R. Altherr, Darrell Ricke, Amanda Ford, Cleo Naranjo, Jason Collins, Michael Lowenstein, Norman Doggett, Larry Deaven and Robert K. Moyzis. Life Sciences Division and Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM.

Individual chromosomes provide the skeletal framework on which genetic data are organized. As the physical map and an ordered assemblage of molecular clones for chromosome 16 neared completion (Doggett, *et al.*, Nature *377 Suppl.*:335), we embarked on the construction of an 'expressed sequence map' of this chromosome. Assuming that there are 100,000 human genes, approximately 3,000 should be encoded by chromosome 16. To this end, we have chosen the strategy of exon amplification to identify expressed sequences on chromosome 16. The strategy employed 96-well plate pools of DNA from the flow sorted and arrayed chromosome 16 cosmid library as the substrate for exon trapping. We have generated and archived more than 3,024 exon clones from 126 plates in the chromosome 16 library. We have sequenced over 2000 of these clones and determined that approximately 50% are distinct. These sequences were analyzed using available databases and are being mapped to specific locations on chromosome 16 using a grided array of previously fingerprinted cosmids that are integrated into a high resolution physical map. We anticipate that this effort will provide a 1000 member exon map of chromosome 16 as the first step toward a chromosome 16 gene map. This effort is complemented by the sample sequencing (SASE) approach to gene identification on chromosome 16 (see Han *et al.* and Ricke *et al.* this meeting).

This work was supported by the Department of Energy under contract W7405-EMG-36.

COSMID BINNING AND cDNA IDENTIFICATION IN HUMAN Xq28

Julia E. Parrish, Evan E. Eichler, Beth A. Firulli, A. Craig Chinault, Mark Graves, Andrew Arenson, Cheng Chi Lee, and David L. Nelson, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030.

The distal long arm of the human X chromosome is one of the most gene-rich regions of the genome; however, of the 22 genetic disorders with Xq28 linkage, 14 remain uncloned. The majority of Xq28 has been assembled into YAC contigs (Palmieri et al., 1994, Genomics 24: 149-158). In order to improve the level of resolution of the physical and transcription maps in this region, we have implemented a two-step strategy utilizing the Lawrence Livermore flow-sorted X cosmid library. YACs are selected to provide 1-3 fold coverage of the region. The YACs are subjected to long-range Alu PCR using six combinations of primers, yielding a range of products up to 13 kb. Products from each PCR reaction are labelled separately, then pooled and hybridized to the cosmid library. Data are entered directly into an in-house database using a graphical interface, which reduces both error and time spent decoding the coordinates of positive signals. Data are then transferred into a relational database, which allows comparison of the results with other screening data generated at Baylor. Cosmids are placed into "bins" defined by hybridization to one or more YACs in the contig. To date, 845 cosmids have been placed into 28 bins in a region of about 4.5 Mb which extends from the FRAXE region to BGN, and 125 cosmids have been placed into 5 bins within the ~2 Mb interval between Factor VIII and the telomere. These efforts provide a substantial degree of coverage of Xq28 in cosmids.

The second step of our approach involves comparison of our data with that generated by Cheng Chi Lee's cDNA/cosmid reciprocal probing strategy. Briefly, his approach is to use pools of individually arrayed cDNA clones (from heart and placental libraries) to probe the cosmid library. Positive cosmids are then used to pursue individual cDNA clones. The pooled cDNA to individual cosmid data are present in the in-house database. Multiple cosmids within a bin which are found to be positive for a given cDNA pool provide evidence for a gene within that bin; moreover, these data allow the cosmids which likely contain that gene to be given first priority in further analysis. We have pursued five such correlations to the single cDNA clone level; additional correlations indicate the presence of two genes, expressed in heart, which lie within the critical region for Barth syndrome.

Supported by grants DE-FG05-92ER61401 and DE-FG03-94ER61830 from the U.S. Department of Energy and a Center grant from the NCHGR of the NIH (NIH 5P30 HG00210) to DLN.

MAPPING AND SEQUENCING OF THE HUMAN X CHROMOSOME

D.L. Nelson[1], E.E. Eichler[1], B.A. Firulli[1], Y. Gu[1], J. Wu[1], E. Brundage[1], A.C. Chinault[1], M. Graves[1], A. Arenson[1], R. Smith[1], E.J. Roth[1], H.Y. Zoghbi[1], Y. Shen[1], M.A. Wentland[1], D.M. Muzny[1], J. Lu[1], K. Timms[1], M. Metzger[1], and R.A. Gibbs[1], [1]Department of Molecular and Human Genetics and Human Genome Center, Baylor College of Medicine, Houston, Texas

The human X chromosome is significant from both medical and evolutionary perspectives. It is the location of several hundred genes involved in human genetic disease, and has maintained synteny among mammals; both of these aspects are due to its role in sex determination and the haploid nature of the chromosome in males. We have addressed the mapping of this chromosome through a number of efforts, ranging from long-range YAC-based mapping to genomic sequence determination.

YAC mapping. The YAC-based map of the X is essentially complete. We have constructed a 40 Mb physical map of the Xp22.3-Xp21.3 region, spanning an interval from the pseudoautosomal boundary (PABX) to the Duchenne muscular dystrophy gene. This region is highly annotated, with 85 breakpoints defining 53 deletion intervals, 175 STSs (20 of which are highly polymorphic), and 19 genes.

Cosmid binning. The YAC-based physical is being used in a systematic effort to identify and sort cosmids prepared at LLNL from flow sorted X chromosomes into intervals. Gene identification through use of a common database for cDNA pool hybridization data is continuing. Additional efforts in distal Xq are described in *Parrish et al.* Over 50 YACs have been utilized as probes to the gridded cosmid arrays. These have identified over 9000 cosmids from the 24,000 member library. An additional 4000 cosmids have been identified using a variety of probes, with the bulk coming from cDNA pool probes.

Cosmid contig construction. Creation of long-range continuity in cosmids proceeds from clones identified by the YAC-based binning experiments. Identification of STS carrying clones is carried out by a combined PCR/hybridization protocol, and adds to the specificity of the overlap data. Cosmids are grown and DNA is prepared by an Autogen robot. DNAs are digested and analyzed by the AB362 GeneScanner for collection of fingerprint data. The use of novel fluorescent dyes (BODIPY) in this application has increased signal strength markedly. End fragment detection is currently carried out with traditional Southern hybridization, however additional dyes will permit detection without hybridization in the GeneScanner protocol. Data are transferred to a Sybase database and analyzed with ODS (J. Arnold, U. Georgia) software for overlap. ODS output is ported to GRAM (LANL) for map construction. A fully automated approach has yet to be achieved, but this goal is increasingly in reach.

Sequencing. An independently funded project awarded to RAG seeks to develop long-range genomic sequence for ~2 Mb of the human X chromosome. In support of this project, cosmids have been constructed and isolated for the 1.6 Mb region between FRAXA and FRAXF in Xq27.3-Xq28. To date, the complete sequences of the regions surrounding the FMR1 and IDS genes have been determined (180 and 130 kb, respectively), along with an additional ~500 kb of the interval. This sequence has led to identification of the gene involved in FRAXE mental retardation. Additional sequence in Xq28 has been determined, including that of a cosmid containing the two genes, DXS1357E and a creatine transporter. This sequence has been duplicated to chromosome 16p11 in recent evolutionary history. Comparative sequence analysis reveals 94% sequence identity over 25 kb, and the presence of pentameric repeats which are likely to have mediated the duplication event. A number of technical advances in sequencing have been developed, including the use of BODIPY dyes in AB373 sequencing protocols, which has offered enhanced base calling due to reduced mobility shifting, improved single strand template protocols for much reduced cost, and streamlined informatics processes for assembly and annotation.

Supported by grants DE-FG05-92ER61401 and DE-FG03-94ER61830 from the U.S. Department of Energy and a Center grant from the NCHGR of the NIH (NIH 5P30 HG00210) to DLN.

A High Resolution FISH Map of Human Chromosome 19 Provides a Metric Backbone For Integration of Physical and Genetic Markers and Assembly of Sequence-Ready Clone Maps

L. A. Gordon, A. Bergmann, M. Christensen, L. Danganan, D. A. Lee, S. Tsujimoto, L. K. Ashworth, A. S. Olsen, A. V. Carrano, H. W. Mohrenweiser and B. F. Brandriff. Human Genome Center, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94550

More than 40 previously unmapped cosmids, many probe-positive for genes and informative genetic markers, have been strategically added to a high resolution fluorescence *in situ* hybridization (FISH) metric map of human chromosome 19 (Brandriff, et al., *Genomics* 23:582-59, 1994; Gordon et al., *Genomics*, in press, 1995); linear order of and distance estimates between neighboring cosmids have been established for 236 cosmids and order is known for over 60 additional cosmids. The current estimate of the combined length of unique sequence portion of both arms of chromosome 19 represented by the map is about 51 Mb with the average interval size between reference cosmids approaching 210 kbp. This map provides a metric backbone that facilitates assembly of BACs, PACs, P1s and cosmids into sequence-ready islands of continuous bacterial-based clones.

The location of new cosmids was determined using standard FISH techniques applied to a series of chromatin targets with increasing resolution. In addition, application of a new technique involving borate swollen interphase nuclei (Yokota et al., *Genomics* 25:485-491, 1995) provided an extended chromatin substrate for three cosmid ordering in two or three colors. The highly extended chromatin in human sperm pronuclei provided confirmation and refinement of order between closely spaced probes and the establishment of distance estimates between cosmids.

The order of and distance between ~150 polymorphic markers (~100 markers with heterozygosities of >0.50) have been determined through inclusion of probe positive cosmids in the FISH metric map. The incorporation of these markers into the physical map resolves the order of tightly linked genetic markers, provides estimates of physical distance between markers and integrates the different linkage maps. The integration of genetic markers into the high resolution FISH metric map and the localization of over 110 genes, combined with approximately 42 Mb of associated EcoR1 restriction maps for which order and distance between islands is known, provides sequence-ready clones that are a unique resource for isolation and sequencing of mapped disease genes.

(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.)

An *EcoRI* restriction map and sequence-ready substrates of human chromosome 19

E. Garcia, A.S. Olsen, L.K. Ashworth, H. Mohrenweiser, M. Burgin, S. Johnson, A.Georgescu, J. M. Elliott, A. Kyle, L. Gordon, T. Slezak, E. Branscomb, and A.V. Carrano. Human Genome Center, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94550

High resolution physical maps of human chromosomes provide the ordered reagents required for detailed analyses of gene organization and furnish the templates for determining the complete DNA sequence. We have developed a cosmid-based ordered physical map that spans approximately 95% of the euchromatin of chromosome 19 (47.5 MB) and that includes complete digest *EcoRI* maps spanning 42 Mb (~83%) of the region. The underlying restriction map defines the minimal number of cosmid clones which are required to sequence the chromosome while minimizing redundant coverage. The present *EcoRI* mapped region is represented by 319 contigs with an average size of 134Kb (range 40-1041 Kb). The clones for selected members of the restriction maps have been anchored to the chromosome by direct FISH or through hybridization to large insert clones that serve as links between the restriction-mapped clones.Thus, the position of each contig is known. Incorporated within the present restriction map are 251 genes, 132 expressed cDNAs, 150 genetic markers and 276 STSs (one STS/180 Kb average).

We will complete the remaining 8 Mb of *EcoRI* restriction coverage of chromosome 19, thereby extending the cosmid coverage of our map, by continuing to convert our available YAC contigs into cosmids. YAC conversion to cosmids is being carried out by interrogating high-density cosmid filter arrays of chromosome 19-specific libraries with Inter-*Alu* PCR and IRS-bubble PCR probes derived from the YACs.

The approach discussed here has enabled the generation of a verified minimum tiling path of cosmids covering greater than 80% of chromosome 19. The present *EcoRI* map includes 76 maps of greater than 150 Kb (average size of 300 Kb).

Current sequencing efforts involve a 1 Mb *EcoRI* mapped region in q13.1 encompassing the CNF gene. Four similarly mapped regions of the chromosome involving the MEF2B gene, OLFR gene family, FHM gene region, and a 2 Mb region including the XRCC1 and ERCC1 DNA repair genes (see poster by Lamerdin et al) are additional foci.

(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.)

IDENTIFICATION, ORGANIZATION AND CHARACTERIZATION OF ZNF GENES IN A 2 MB CLUSTER ON 19p12

Evan E. Eichler [1], *Susan Hoffman and Harvey Mohrenweiser.* Lawrence Livermore National Labs, Livermore, CA 94550.

Zinc finger (ZNF) genes represent one of the largest and most diverse family of genes found in the human genome. These genes encode transcriptional regulators which are believed to play critical roles in cellular and developmental differentiation processes. DNA binding of the encoded proteins is typically mediated by a zinc finger motif which consists either of two cysteines and two histidines (Krüppel family or C_2/H_2 type) or four cysteines alone (steroid receptor or C_2/C_2 type). It has been estimated that there are between 300-700 loci in the human genome which show homology to the C_2/H_2-type zinc finger motif. Despite their abundance, the function of the vast majority of these genes is not known. Chromosome 19 appears to be particularly enriched for zinc finger genes such that one third of all ZNF loci are distributed within three potential clusters corresponding to cytogenetic band locations 19p12, p13.2 and q34.

We are currently working to map and characterize all ZNF genes in the 19p12 cluster between STS markers D19S269 and D19S450. Our strategy has been to develop a highly integrated physical map of this region using FISH, STS markers and conversion of overlapping YAC contigs to a cosmid clone map of this interval. Eight overlapping YAC clones have been identified in the 19p12 region, creating a contig of approximately 2 megabases. YAC DNA was purified and used to screen a flow-sorted chromosome 19 arrayed cosmid library. A total of 200 cosmids were identified and assembled into contigs based on fluorescent fingerprinting methods. The location of cosmid contigs was confirmed using FISH and cosmids were anchored using a STS marker screening strategy. Sixteen cosmid contigs have been constructed comprising approximately 1.4 MB of this 2MB interval. In addition a detailed *EcoR* I enzyme restriction map has been developed for approximately 650 kb of this region. As a first attempt of interdigitating potential ZNF genes in this region, cosmids were screened with a degenerative 27-mer oligonucleotide probe to the conserved "H/C" link between zinc finger motifs. This analysis identified a minimum of 12 potential C_2/H_2 ZNF genes in this region. Hybridization and sequence analysis has confirmed the identity, thus far, of one functional zinc finger gene (ZNF 85).

Once the physical map is completed, cDNAs will be identified and assigned to an *EcoR* 1 restriction enzyme map of this cosmid contig. The development of a high-resolution transcription map, followed by the determination of expression profiles of various ZNF cDNAs should provide valuable insight into the functional properties of active ZNF genes and their potential role in development and differentiation. Furthermore, the elucidation of the organization of this cluster by novel Alu-typing strategies as well as subfamily assignment will allow us to test models of gene duplication and may suggest other mechanisms of gene evolution for 19p12 and other ZNF clusters in the human genome.

[1]DOE Human Genome Distinguished Postdoctoral Fellow

The Application of Exon Trapping, cDNA Direct Selection, Directed Genomic Sequencing and Computational Genomics to the Discovery of Genes in a 20q13.2 Breast Cancer Amplicon.

Colin Collins[1], *Soo-in Hwang*[1], *Johanna Rommens*[2], *David Kowbel*[1], *Christopher Martin*[1], *Michael Palazzolo*[1], *Gordon Hutchinson*[3], *Tony Godfrey*[4] *and Joe W. Gray*[1,4].

[1]Lawrence Berkeley National Laboratory, Berkeley, California, [2]Hospital for Sick Children, Toronto, Ontario, [3]RabbitHutch Biotechnology, 100 Mile House, British Columbia, [4]Division of Molecular Cytometry, University of California, San Francisco, California.

In developed and several developing countries breast cancer is one of the most frequently diagnosed neoplasms and the leading cause of cancer related death amongst women. The mortality rate for breast cancer is approximately 27 per 100,000 women. Pangenomic surveys using comparative genomic hybridization (CGH) and fluorescence *in situ* hybridization (FISH) have revealed >20 regions of allelic imbalance suggesting the presence of numerous previously unrecognized tumor suppressor genes and oncogenes. Evidence is accumulating that allelic imbalance at these loci may play an important role in neoplastic transformation, progression and development of resistance to chemotherapeutic agents.

Chromosome 20 band q13.2 is amplified in 40% of breast cancer cell lines and 29% of primary breast tumors. Moreover, high level amplification (7% of primary tumors) has been shown to be associated with decreased disease-free survival and an increased S-phase fraction. We have cloned the 20q13.2 amplicon as a 2 Mb sequence-ready P1 and BAC contig and localized the minimum common region of amplification to a ~600 kb interval by performing interphase FISH on primary tumors with 9 P1 probes spanning the contig. To identify genes in the amplicon that program the aggressive phenotype of these breast tumors we are: (1) applying exon trapping and cDNA direct selection to the P1 and BAC contig and (2) in collaboration with the LBNL Human Genome Center (HGC) sequencing the 600 kb amplicon.

Exon trapping has been performed on the P1 and BAC clones spanning the minimum common region of amplification by digesting the clones with Pst1 and Sac1 followed by subcloning into the pSPL3 exon trapping vector. To date, >30 exons have been isolated and sequenced from the 600 kb interval. Computational analysis using BLAST has revealed homologies to known genes, ESTs and *S. cerevisiae* chromosome XIV. cDNA direct selection has been performed with P1 and BAC clones spanning the amplicon using pooled cDNA synthesized from 9 tissues and cDNA from the breast cancer cell line BT474. Preliminary studies have localized four cDNAs to the core of the amplicon.

Directed genomic sequencing is being employed to sequence the minimum common region of amplification. Presently, three contiguous P1 clones spanning approximately 200 kb are being sequenced. Putative exons are being identified in the genomic sequence using the programs GRAIL2, XGRAIL, SORFIND and BLAST. Exons identified by exon trapping and genomic sequencing are expanded by performing RACE-PCR and analyzed for phylogenetic conservation by hybridization to zoo blots. Exons and cDNAs are being assessed for expression by performing RT-PCR and Northern hybridization using RNA isolated from appropriate breast cancer cell lines and primary tumors. It is expected that the combination of exon trapping, direct selection, and directed genomic sequencing will culminate in the complete molecular description of the 20q13.2 amplicon resulting in the identification of the hypothesized oncogene(s), improved diagnosis and prognostication, molecular therapeutics and ultimately decreased mortality in breast cancer.

This work was supported by grants from US DOE contract DEAC0376SF00098, USPHS grants CA44768, CA45919, CA52807 and Vysis.SH is supported the Human Genome Distinguished Postdoctoral Fellowship from DOE/ORISE.

ASSEMBLING A LINKAGE MAP FOR THE DOG GENOME

Mark W. Neff[1,2], Mike Strathmann[1], Janet Ziegle[3], and Jasper Rine[1]. [1]Department of Molecular & Cell Biology, Division of Genetics, University of California, Berkeley, CA 94720. [2]Human Genome Center, Lawrence Berkeley National Laboratory, Berkeley, CA 94720. [3]Applied Biosystems (ABI), Foster City, CA 94404. email: neff@mendel.berkeley.edu

The morphological and behavioral traits that distinguish breeds of domestic dog are genetically defined, having been fixed by artificial selection. A linkage map of the dog genome will make breed-specific traits and canine genetic diseases amenable to mapping approaches. We and others are assembling a linkage map for the domestic dog using microsatellites, molecular markers which are highly polymorphic and relatively straightforward to genotype [1,2]. We currently genotype these markers with an ABI fluorescence-based DNA sequencing instrument. Multiplex marker sets for dog have been established based upon differences in product size and primer dye color, allowing for high throughput analysis and unambiguous allele assignment [3].

We have also investigated whether alternative types of genetic markers might supersede microsatellites, with the aim of reducing the time and cost of constructing and using linkage maps. Specifically, an alternative marker should obviate the need for DNA sequencing and primer synthesis during marker construction, and should permit a single mass genotyping of the collection of markers for each individual. RFLPs have the potential to meet these requirements. Bi-allelic RFLPs are sufficiently informative for an F_2 intercross when the parents are homozygous for different alleles. To create a library of RFLPs that are informative for a Border Collie X Newfoundland intercross, we have begun a series genomic subtractions using DNA from the parents [4]. We have also begun developing novel methods of RFLP genotyping which do not require DNA sequence information or marker-specific primers. If successful, these advances will reduce the resources necessary to construct and use linkage maps. A low cost method of genotyping will permit genetic mapping in any species that possesses interesting natural variation.

M.W.N. was supported by a DOE Human Genome Distinguished Postdoctoral Fellowship administered by ORISE, and M.S. was supported by an Alexander Hollander Postdoctoral Fellowship.

[1] E. A. Ostrander, G. F. Sprague, and J. Rine (1993) Identification and characterization of dinucleotide repeat (CA)n markers for genetic mapping in dog. Genomics 16:207-13.
[2] E. A. Ostrander, F. A. Mapa, M. Yee, and J. Rine (1995) One hundred and one new simple sequence repeat-based markers for the canine genome. Mammalian Genome 6:192-5.
[3] J. S. Ziegle et al. (1992) Application of automated DNA sizing technology for genotyping microsatellite loci. Genomics 14:1026-31.
[4] M. Rosenberg, M. Przybylska and D. Straus (1994) "RFLP subtraction": a method for making libraries of polymorphic markers. Proc. Natl. Acad. Sci. USA 91: 6113-7.

Structural and Functional Analysis of a Conserved Zinc-Finger Gene Cluster in Man and Mouse

Mark Shannon[1,3], Linda Ashworth[2], Jane Lamerdin[2], Loren Hauser[1], Elbert Branscomb[2], and Lisa Stubbs[1]

[1]Biology Division, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, TN 37831-8080; [2]Human Genome Center, Lawrence Livermore National Laboratory; [3]Corresponding author

As part of a comprehensive man-mouse comparative mapping study of human chromosome 19 being conducted as a collaboration between the Biology Division at the Oak Ridge National Laboratory and the Human Genome Center at the Lawrence Livermore National Laboratory, we have previously identified a conserved cluster of Kruppel-type zinc-finger (ZNF) genes distal to *XRCC1* in 19q13.2 and the syntenically homologous region of mouse chromosome 7. Here we report current findings from an ongoing study of the structure, function, and evolution of this apparently homologous pair of gene clusters. Our results indicate that each cluster consists of at least ten related Kruppel-associated box (KRAB)-containing ZNF genes and that the human genes are arranged in tandem over a distance of 350-450kb. We have also found that the KRAB A domains associated with the mouse and human *XRCC1*-linked ZNF gene clusters are highly similar and are clearly distinct from those of ZNF genes located elsewhere in either genome. Several cDNA clones representing genes in the murine cluster have been isolated using the cluster-identifier 19q13.2 KRAB sequence (hkraba1) as a probe, and three clones have been analyzed in detail. The KRAB A domains of these genes are nearly identical, but other portions of the genes, including the DNA-binding ZNF domains, differ considerably. Interestingly, despite the divergence of ZNF sequences during elaboration of the cluster, Southern analysis suggests that these portions of orthologous mouse and human genes have been remarkably well conserved. Taken together, these data suggest that this cluster of genes may have evolved to encode proteins that recognize different sites in DNA as a result of variant ZNF sequences, while interacting with a common set of transcription factors due to highly similar KRAB domains. Finally, a survey of expression has shown that the mouse genes have highly specific, partially overlapping, but clearly distinct patterns of expression in adult tissues. Transcription factors of the Kruppel-type ZNF subclass presumptively make up nearly 1% of all genes in mammals, are widely found in such contiguous clusters, and are about 16-fold over-represented on human chromosome 19. Our studies open the way for a systematic structural, functional, and evolutionary analysis of these clusters and their constituent genes.

This work was supported by USDOE under contract DE-AC0584OR21400 with Lockheed-Martin Energy Systems, Inc., and contract W-7405-ENG-48 with the Lawrence Livermore National Laboratory. M.S. is supported by a DOE Human Genome Distinguished Postdoctoral Fellowship.

GENE DISCOVERY ON HOMOLOGOUS MOUSE AND HUMAN COSMIDS CONTAINING ZNF GENES.

Loren Hauser, Mark Shannon, Melissa York, Lisa Stubbs, and Richard Mural. Biology Division, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, TN 37831-8080

One of the most efficient ways of finding genes on cosmid or P1 clones is to shotgun sequence random fragments and analyze the data for potential coding regions using computer programs such as GRAIL. In gene rich regions of the mouse or human genome, a cosmid sized clone (about 40kb) may contain 3 to 5 genes. In such regions about 1 in 5 to 1 in 10 random 1kb subclones should contain protein coding exons and over 90% of these will be recognized by GRAIL analysis. A joint comparative human-mouse mapping project between Lawrence Livermore National Laboratory and Oak Ridge National Laboratory has previously identified a conserved cluster a Kruppel-type zinc-finger (ZNF) genes distal to XRCC1. Mapping data using mouse KRAB and zinc-finger probes placed one ZNF gene about every 25 kb in a tandem arrangement on the human cosmid contig. Therefore, there should be 1-2 ZNF genes per cosmid. A probe from a unique region of a cDNA mapped to this cluster was used to identify a cosmid from both a mouse and human library. Shotgun libraries from each cosmid were prepared in M13 and sequenced. The data was analyzed directly using batchGRAIL. BatchGRAIL is designed to search for potential coding regions and if the coding potential exceeds a threshold its translation is searched against Swissprot. Consistent with expectation, approximately 1 in 10 clones are positive for ZNF coding regions. In addition, a LINE was identified via its weak homology to DNA polymerases. Comparative sequence of the mouse and human cosmids will be presented. The structure of both human and mouse cosmids containing orthologous genes from this cluster will provide useful information on the evolution and possible function of the cluster.

This Research was supported by the Office of Health and Environmental Research, United states Department of Energy, under Contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.

CONSTRUCTION OF MICE WITH LARGE DELETIONS IN THE INTERLEUKIN GENE CLUSTER FOR FUNCTIONAL ANALYSES OF HUMAN GENES IN THE SYNTENIC REGION

Yiwen Zhu, Miles Miller, Jan-Fang Cheng, and Eddy Rubin
Human Genome Center, Life Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

The interleukin gene cluster on human chromosome 5q31 was selected for large scale sequencing because it contains a number of functionally related genes and is therefore likely to yield novel cytokine gene candidates through sequence analysis. Parallel to sequencing and sequence analysis, we have begun to develop a mouse model to study functions of the sequenced genes.

Saturated mutation of a defined region in the mouse genome has only recently become feasible with the very recent development of a new genetic engineering technique, the Cre-Lox site-specific recombination. It is based on the ability of Cre recombinase to delete regions bound by loxP sites (a 34-bp sequence) *in vivo*. Application of such technology to mouse embryonic stem (ES) cells would allow large chromosome deletions to be introduced into the germline and therefore would facilitate new functional analyses of the genome. The size of deletions could be multimegabases since translocation between mouse chromosomes has been previously reported for the Cre-Lox recombination. Multimegabase deletions, however, frequently produce an embryonic lethal phenotype. One key advantage of the Cre-Lox approach is the ability to generate deletions specific to a developmental stage and tissue. LoxP targeted ES cells in which the deletion has not yet been activated would be used to achieve germline transmission. The deletion could then be induced by crossing ES cell-derived progeny with transgenic strains in which Cre recombinase expression is under control of a developmental stage-specific and tissue-specific promotor.

In our targeting experiments, loxP sites were inserted into the murine homologue of the human 5q31 region. The availability of previously cloned and sequenced genes (IL4, IL13, IL5, IRFI, CSF2, IL3) dispersed in the targeted region of mouse chromosome 11 provided the required sequence information for creating targeting vectors. Four targeting vectors carrying loxP sequences and either neomycin (Neo^r) or hygromycin (Hyg^r) markers were designed to integrate into 4 locations in the gene cluster. These locations are (1) 5' flank of the IL13 gene, (2) 10.5 Kb upstream of the IL5 gene, (3) internal of the IRF1 gene, and (4) 3' flank of the CSF2 gene. All 4 targeting vectors were successfully inserted into the predetermined sites. We have confirmed the targeted integrations using restriction digests and Southern blot analysis of the ES cell genomic DNA. The frequency of appropriated targeting event varied from 46% to 4%.

We are now in the process of injecting targeted ES cells into mouse embryos. This will allow the production of mouse strains in which all the cells contain the targeted DNA. These animals will be bred with animals expressing Cre recombinase in either all or selected tissues (such as lymphoid) to create a series of nested deletions. The progeny of these matings will be characterized phenotypically to identify specific sequences whose absence impacts on the development of certain cell lineages in the immune system. This *in vivo* analysis combined with 1.2 Mb of sequence data from human 5q31 will contribute to the discovery and functional analysis of genes contained within this targeted region of the human genome.

Identification of genes affecting learning and memory in chromosome 21 YAC transgenic mice

Desmond J. Smith, Mary E. Stevens, Jan-Fang Cheng and Edward M. Rubin

Human Genome Center, Lawrence Berkeley National Laboratory, University of California, 1 Cyclotron Road, Berkeley, CA 94720

Moving from genomic sequence information to an understanding of biological function is a problem that is beginning to loom large as the genome project moves forward. As a pilot investigation of this problem, we have created a 2 Mb *in vivo* library from the Down syndrome region of human chromosome 21q22.2 in transpolygenic mice in order to link defined DNA sequences and genes to biological function.[1] Mice were created using 4 YACs (230E8, 141G6, 152F7 and 285E6, varying between 400-700 kb) and 4 P1s from human chromosome 21q22.2. The mice were screened for new genes based on the phenotypic effects on the host organism.

Our assays especially concentrate on behavioral traits such as learning and memory, since defects in these processes are the most important consequences of an extra copy of the Down syndrome region in humans. Learning and memory in the transpolygenic mice was evaluated using the Morris water maze test and the activity of the mice is being evaluated using an open field test. Mice bearing YACs 141G6 and 285E6 did not differ from non-transgenic controls. The two separate founder lines of mice created bearing YAC 152F7 display deficiencies in learning suggestive of specific hippocampal abnormalities and also show abnormalities on an open field test. Both lines of mice containing the YAC 152F7 also display a phenotype which we call "*Clencher* ", clasping of all four limbs together when the animals are suspended by the tail. The process of microinjection in addition to producing mice which contain the intact YACs also produces mice that contain fragments of the YAC. We are now examining four different lines of mice containing different segments of YAC 152F7 and preliminary studies suggest that one of these lines shares the phenotype seen in mice containing the intact YAC while the others appear normal. We hope that molecular characterization of these different lines can be used to localize the gene on YAC 152F7 responsible for the phenotype seen in animals containing the intact YAC.

Mice harboring YAC 230E8 demonstrate very distinct abnormalities from that observed in the YAC 152F7 transgenics. The YAC 230E8 transgenics are profoundly deficient in their performance on the Morris maze, indicating functional abnormalities abnormalities throughout the cerebral cortex. These studies thus demonstrate that two distinct segments of DNA from the Down syndrome region can cause distinct phenotypes of abnormal learning and memory when expressed in transgenic mice. The approaches described in this abstract can be used to investigate other regions of the human genome as a general strategy for linking gene discovery and phenotype.

[1] D.J. Smith et al., *Genomics* **27**, 425-434 (1995).

Comparative Mapping of Mouse Chromosome 13

Charles M. Perou[1], *Antoine L. Perchellet*[2], *Jerry Kaplan*[1], *and Monica J. Justice*[3],
1 Department of Pathology, University of Utah School of Medicine, Salt Lake City, Utah 84132, 2 Division of Biology, Kansas State University, Manhattan, KS 66506, and 3 Biology Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831

Mouse Chromosome 13 contains regions conserved on human chromosomes 1q41-q43, 6p23-p21, 7p22-p13 and 5q11.2-q35. These regions contain numerous models for human disease, such as Chediak Higashi syndrome (mouse *beige*) [1], Greig cephalopolysyndactly syndrome (mouse *extra-toes*) [2], and cancer [3, 4, 5, 6]. The mouse provides a useful genetic model organism for understanding the mechanism of disease in humans, as well as for dissecting the biological function of other genes that are conserved between mouse and human.

As a first step in examining the function of genes in the proximal region of mouse chromosome 13, we are creating a fine structure genetic linkage map of the genomic region encompassing *beige* (*bg*) and *satin* (*sa*). An interspecific backcross involving SB/Le and *Mus spretus* mice was used to generate a molecular genetic linkage map of mouse chromosome 13 [7]. This map provides the gene order of the two phenotypic markers *bg* and *sa* relative to restriction fragment length variants and simple sequence length variants. Our initial study involves 132 backcross animals, and spans the entire chromosome. The results from these data will direct "interval mapping" of the *bg-sa* region on an additional 400 animals. In the *bg-sa* region of 6 cM, the marker density of our map exceeds 7 markers/cM.

In parallel to the genetic linkage mapping, we are creating a physical map of the region using *Nidogen* (*Nid*) as a molecular starting point for cloning a YAC contig. Multiple cDNAs have been isolated from these YACs, and the results show that linkage homology with human chromosome 1q41-43 is highly conserved in the mouse. The results also precisely localize a breakpoint in homology between human chromosomes 1q43 and 7p13 in the region. The genetic and physical mapping results will provide valuable resources for further functional studies of the conserved genes in the region using induced mutations.

MJJ is supported by the U.S. Department of Energy, work proposal No. ERKP057, by the National Institutes of Health, 7R29CA63229-02, and by an award from the American Cancer Society, JFRA-553. J.K. is supported by the National Institutes of Health, HL26922, and C.M.P. is the recipient of an NIH Genetics Training Grant T32GM07464.

1. Holcombe, R.F., Strauss, W., Owen, F.L., Boxer, L.A., Warren, R.W., Conley, M.E., Ferrara, J., Leavitt, R.Y., Fauci, A.S., Taylor, B.A., and Seidman, J.G. (1987) *Genomics* **1**, 287-291
2. Winter, R.M. (1988) *J Med Genet* **25**, 480-487
3. Rousseau-Merck, M., Bernheim, A., Chardin, P., Miglierina, R., Tavitian, A., and Berger, R. (1988) *Hum Genet* **79**, 132-136
4. Van Cong, N., Fichelson, S., Gross, M.S., Sola, B., Bordereaux, D., de Tand, M.F., Guilhot, S., Gisselbrecht, S., Frezal, J., and Tambourin, P. (1989) *Hum Genet* **81**, 257-263
5. Vortkamp, A., Franz, T., Gessler, M., and Grzeschik, K. (1992) *Mammalian Genome* **3**, 461-463
6. Hsieh, C., Vogel, U.S., Dixon, R.A.F., and Francke, U. (1989) *Somatic Cell Mol Genet* **15**, 579-590
7. Justice, M.J., Silan, C.M., Ceci, J.D., Buchberg, A.M., Copeland, N.G., and Jenkins, N.A. (1990) *Genomics* **6**, 341-351

COMPARATIVE APPROACHES TO THE ANALYSIS OF HOMOLOGOUS MOUSE AND HUMAN GENOMIC REGIONS

[1]*Lisa Stubbs*, [1]*Johannah Doyle*, [1]*Ethan Carver*, [1]*Karen Rollins*, [1]*Laura Chittenden*, [1]*Mark Shannon*, [1]*Joomyeong Kim*, [2]*Linda Ashworth, and* [2]*Elbert Branscomb*, [1]Biology Division, Oak Ridge National Laboratory, P.O.Box 2009, Oak Ridge, TN 37831-8077; and [2]Human Genome Center, P.O. Box 808, Lawrence Livermore National Laboratory, Livermore, CA 94550.

Numerous studies have confirmed the notion that mouse and human chromosomes resemble each other closely within blocks of syntenic homology that vary in widely size, containing from just a few to several hundred related genes. Within the best-mapped of these homologous regions, the presence and location of specific genes can be accurately predicted in one species, based upon the mapping results obtained in the other. In addition, information regarding gene function derived from the analysis of human hereditary traits or mapped murine mutations, can also be extrapolated from one species to another. However, syntenic relationships are still not established for many human regions, and local rearrangements including apparent deletions, inversions, insertions, and transposition events, complicate most of the syntenically homologous regions that appear simple on the gross genetic level. Because of these complications, the power of prediction afforded in any homology region increases tremendously with the level of resolution and degree of internal consistency associated with a particular set of comparative mapping data. Our groups have been interested in further defining the borders of syntenic linkage groups in human and mouse, and upon devising means of exploiting the relationships between the two genomes for the discovery of new genes and other functional units in both species.

One of the larger contiguous blocks of mouse-human genomic homology includes the proximal portion of mouse chromosome 7 (Mmu7), and all murine homologs of genes mapping to human chromosome 19q (H19q) that have been recorded to date. Detailed analysis of this large region of mouse-human homology have served as the initial focus of these collaborative studies. Our results have shown that gene content, order and spacing are remarkably well-conserved throughout the length of this approximately 23 cM/29 Mb region of mouse-human homology, except for (1) an overall "inversion" of sequences relative to the centromere, (2) three apparent "transpositions" of gene-rich segments in mouse relative to man, and (3) two inversions involving smaller subregions. One of these differences involve a small segment of H19q13.4 genes whose murine counterparts have been transposed out of the large Mmu7/H19q conserved synteny region into a separate linkage group located on mouse chromosome 17. The five internal rearrangements are clustered together at two sites, suggesting either the coincidence of rearrangement events or their common association with unstable DNA sequences. Interestingly, both rearranged regions are occupied by large tandemly clustered gene families, suggesting that these locally repeated sequences may have contributed to their evolutionary instability. More recently, we have extended mapping studies to include other regions, and are working to define the borders of mouse-human syntenic segments on a broader, genome-wide scale.

As another aspect of this collaborative project, we have explored means of exploiting mouse-human genomic conservation in the isolation of functionally-significant sequences from large cloned regions of human DNA. Using a model system comprised of sequenced mouse and human cosmids spanning the *XRCC1* gene, we have succeeded in developing a method for isolating exons as well as conserved regulatory sequences with high efficiency. We have recently applied this methodology to the analysis of two larger genomic regions (~300 kb each) spanned by overlapping human cosmids and parallel sets of mouse P1 clones. Our analysis of clones isolated from these two regions, which include several known genes as well as large segments of unexplored DNA, will be presented. Because of its relative simplicity and ability to purify both coding and regulatory regions at high yield, this conserved-element purification method holds great promise as an efficient tool for gene discovery in cloned genomic regions.

This work was supported by USDOE under contract DE-AC0584OR21400 with Lockheed-Martin Energy Systems, Inc., and contract W-7405-ENG-48 with the Lawrence Livermore National Laboratory.

MOUSE TRANSLOCATION MUTANTS PROVIDING MODELS FOR A WIDE VARIETY OF RELATED HUMAN HEALTH DEFECTS

[1]*Lisa Stubbs*, [1]*Cymbeline Culiat*, [1]*Ethan Carver*, [1]*Nestor Cacheiro*, [2]*Gary Wright, and* [1]*Walderico Generoso*, [1]Biology Division, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, TN, 37831-8077; and [2]University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Blvd., Dallas, TX 75235.

Balanced translocations have proved invaluable tools in the mapping and molecular cloning of a number of different acquired and inherited human diseases including Neurofibromatoma type I, Autosomal recessive polycystic kidney disease, and several different types of human cancers. Because the balanced translocations are cytologically visible, and generally produce profound disturbances in both gene expression and DNA structure, this type of mutation provides a valuable "tag" that greatly simplifies mapping, cloning, and assessment of candidate genes associated with a disease. Although balanced translocations are relatively rare in human populations, they are readily induced in the mouse. Using various mutagenesis protocols, we have generated numerous translocation-bearing mutant mouse strains that display an impressive variety of health-related anomolies, including limb and skeletal deformities, neural tube defects, ataxias, tremors, hereditary deafness and blindness, reproductive dysfunction, and complex behavioral defects. The ability to map the genes associated with translocation breakpoints cytogenetically, first crudely through straightforward banding techniques and then to a higher level of resolution using fluorescence *in situ* hybridization methods, allows us to avoid the costly and time-consuming crosses that are required for the mapping of most mutant genes. With this rapid, crude-level mapping information, we can readily assess possible relationships between newly arising mutant phenotypes and linked candidate genes, or related diseases that map to the homologous regions of specific human chromosomes. Using this approach, we have recently begun to define the map positions of several mutations, including one producing hydrocephalus, two associated with progressive ataxia, one causing a subtle and complex behavioral disorder, and two producing different types of congenital inner ear defects. Mapping results have clearly indicated that one of these mutations affects the murine homolog of the gene disrupted in Usher's syndrome type 1C, an uncloned human disorder associated with severe congenital deafness, balance defects, and early-onset blindness. We have recently identified a candidate gene corresponding to this mutation, and are poised to begin investigations of the gene's role in inner ear and eye development in mice and in affected human families.

To date, we have characterized and mapped only a fraction of the large and growing number of translocation-bearing strains that comprise this valuable mutant collection. The breakpoints that have been mapped so far, which represent primarily those associated with severe, early onset, and easily detectable phenotypes, are scattered widely throughout the mouse genome and map to a broad selection of human homology regions. Over the next few years, as new breakpoints are located and large numbers of newly-sequenced cDNA clones are assigned to the mouse and human maps, the potential for rapid association between cloned gene and mapped mutation will no doubt increase dramatically. This large collection of murine translocation mutants therefore represents a powerful resource for linking mapped cDNA clones to health-related phenotypes throughout the genome.

This work was supported by USDOE under contract DE-AC0584OR21400 with Lockheed-Martin Energy Systems, Inc.

Conservation of a Zinc-Finger Gene Cluster, ZNF 134-like, in Man and Mouse

Joomyeong Kim[1,3], Mark Shannon[1], Linda Ashworth[2], Elbert Branscomb[2], and Lisa Stubbs[1]

[1]Biology Division, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, TN 37831-8080; [2]Human Genome Center, Lawrence Livermore National Laboratory; [3]Corresponding author

Several lines of evidence now suggest that the human genome carries hundreds of zinc-finger (ZNF)-containing genes and that many of these genes are arranged in clusters. Curiously, a disproportionate number of these genes map to human chromosome 19, and detailed physical mapping performed at Lawrence Livermore National Laboratory's Human Genome Center has further demonstrated that ZNF genes are primarily located in six major clusters dispersed throughout the length of the chromosome. As part of an extensive man-mouse comparative mapping study of human chromosome 19 conducted by investigators at the Oak Ridge and Lawrence Livermore National Laboratories we have targeted for in-depth study a region spanning approximately 2.5Mb near the telomere of H19q and occupying most of subband 19q13.4. Since previous studies have indicated that genes bordering this region are conserved in proximal mouse chromosome 7 (Mmu7), we predicted that the large array of ZNF genes known to be present in 19q13.4 would be similarly conserved in the syntenically homologous region of Mmu7.

As a preliminary step for characterization of this region, we have assigned several known ZNF cDNA sequences to the 19q13.4 physical map. Two genes, ZNF134 and ZNF132, have been localized to centrally located contigs (577/1514 and 32, respectively) using probes under high stringency conditions. However, under low stringency conditions, these two probes detected overlapping sets of contigs, which include several neighboring 19q13.4 contigs, suggesting that the two genes are members of a large clustered family. Physical mapping of several ZNF-positive contigs from this central region suggests that each carries 5 or more related genes and that the entire region is likely to contain 50 or more. We have also used these gene probes to map related sequences in the mouse using the interspecific backcross system. As expected, each probe detected several loci, which mapped together in the related region of Mmu7. These data suggest that the mouse genome contains a ZNF134-related gene cluster in a region that is syntenically homologous to human chromosome 19q13.4. These largely unexplored regions provide a rich resource for studying the structure, function, and evolution of the many clustered gene families located throughout the human and mouse genomes.

This work was supported by USDOE under contract DE-AC0584OR21400 with Lockheed-Martin Energy Systems, Inc., and contract W-7405-ENG-48 with the Lawrence Livermore National Laboratory.

GENETIC BAR CODING: DETECTION AND LOCALIZATION OF SEQUENCE CHANGES USING CLEAVASE™ FRAGMENT LENGTH POLYMORPHISM (CFLP™) ANALYSIS

Mary Ann D. Brow, Mary C. Oldenburg*, Victor Lyamichev*, Laura M. Heisler*, Natasha Lyamicheva*, Jeffrey Groteluechen*, Richard Handrow*, Sergei Kozyavkin*, D. Michael Olive*, Lance Fors*, Lloyd M. Smith♣ and James E. Dahlberg♠*
*Third Wave Technologies, Inc., 2800 S. Fish Hatchery Rd., Madison, WI 53711
♣University of Wisconsin, - Madison, Department of Chemistry
♠University of Wisconsin, - Madison, Department of Biomolecular Chemistry

Methods for the detection of genetic mutations resulting in human diseases such as cancer or infection with drug resistant *Mycobacterium tuberculosis* have become a critical need for effective healthcare. While some of these mutations may cause significant changes in the responsible genes, many of these mutations, such as those found associated with the p53 gene, are changes in single nucleotides. Methods such as single strand conformation polymorphism (SSCP) and denaturing gradient gel electrophoresis (DGGE) have yielded limited success in detecting mutations in short DNA fragments. Under the best conditions, these methods are limited to indicating the presence or absence of base changes, some of which may not result in phenotypic change. Until now, DNA sequencing has been the only method which can successfully identify phenotypically significant mutations, yet the cost and complexity of DNA sequencing has prevented its general use by the clinical community. We report here the use of Cleavase™, a thermostable structure-specific endonuclease, for detection of mutations in clinically significant genes including b-globin, p53, and the *M. tuberculosis rpo*B and *kat*G genes. Cleavase™ is capable of recognizing and cutting secondary structures formed in the single strands of a DNA fragment following denaturation by heating and subsequent cooling. The digested single stranded DNA fragments are rapidly resolved by electrophoresis on denaturing acrylamide gels followed by detection by means of either a radioactive or a non-radioactive label. The presence of a mutation is indicated by a change in the fragment pattern near the region of the mutation. Thus the Cleavase™ reaction, referred to as Cleavase Fragment length Polymorphisms (CFLP), can be used for both detection and localization of mutations. We have been able to detect 102 of 103 polymorphisms in a total of 15 genetic systems. Point mutations in cDNA clones spanning exons 5 through 8 of the human p53 gene were reproducibly detected and differentiated. Using CFLP analysis, mutations resulting in a drug resistance phenotype could be reliably distinguished from phenotypically silet mutations in *M. tuberculosis* isolates. The simplicity, and rapidity of CFLP should facilitate the use of mutation analysis as a routine technique in the analysis of human disease processes.

Supported by a grant from the Department of Commerce, National Institutes of Standards and Technology Advanced Technology Program under Proposal Number 94-05-0012

Triplet repeat fingerprinting: enhanced methods for comparative genome analysis

Charles R. Cantor, Natasha Broude, Ronald Yaar, and Cassandra L. Smith
Center for Advanced Biotechnology, Departments of Biomedical Engineering, Pharmacology, and Biology, Boston University, Boston MA 02215

Triplet repeats like $(GGC)_n$ are an important class of human genetic markers, and they are also responsible for a number of inherited diseases involving the central nervous system. For both of these reasons it would be very useful to have a way to monitor the status of large numbers of triplet repeats simultaneously. We are developing methods to isolate and profile classes of such repeats.

In one method, genomic DNA is cut with one or more restiction nucleases, and splints are ligated onto the ends of the fragments. Then fragments containing a specific class of repeats are isolated by capture on magnetic microbeads containing an immobilized simple repeating sequence. The desired material is then released, and, if necessary, a selective PCR is done to reduce the complexity of the sample. Otherwise the entire captured sample is amplified by PCR. The spectrum of repeats is then examined by electrophoresis on an automated fluorescent gel reader. In our case the Pharmacia ALF is used because of its excellent quantitative signal accuracy. A very complex spectrum of bands is seen representing hundreds of DNA fragments. We have shown that this spectrum is dramatically different with DNAs from unrelated individuals, and the spectrum is markedly dependent on the choice of restriction enzyme, as expected. Repeated measurements on the same sample are highly reproducible. The ability of the method to detect a specific altered repeat length in a complex DNA sample has been validated by examining several individuals with normal or expanded repeat sequences in the Huntington's disease gene. One very powerful application of this method may be the analysis of potential DNA differences in monozygotic twins discordant for a genetic disease. This method can be used to capture genome subsets containing any interspersed repeat. It will also detect insertions and deletions nearby such repeats. Methylation differences between sensitive methylation samples are also detectable when restriction fragments are used.

Conventional analysis of triplet repeats is very laborious since individual repeats must be analyzed by electrophoresis on DNA sequencing gels. The decrease in effort for such analyses will scale linearly as the number of repeats that can be analyzed simultaneously, so we are potentially looking at something like a factor of 100 improvement if the above scheme under development can be effectively realized.

As an alternative approach, we are developing chip-based methods that can detect the length of a tandemly-repeating sequence without any need for gel electrophoresis. Here the goal is to build an array of all possible repeat sequence lengths flanked by single-copy DNA. When an actual sample is hybridized to such an array, the specific alleles in the sample will produce perfect duplexes at their corresponding points in the array and at mismatched duplexes elsewhere. Thus, the task of scoring the repeat lengths is reduced to the task of distinguishing perfect and imperfect duplexes. Currently we are exploring a number of different enzymatic protocols that offer the promise of making such distinctions reliably.

High Resolution Physical Mapping of EcoRI restriction Sites on Intact Cosmids by AFM Imaging

D. P. Allison, P. S. Kerper, M. J. Doktycz, T. Thundat, and R. J. Warmack
*Health Sciences Research Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6123

Physical mapping of genomic DNA at 100 base resolution using direct AFM imaging has been accomplished. This new mapping technology was created by attaching a mutant E*co*RI endonuclease[(1)], deficient in cleavage function, to double-stranded DNA and imaging the attachment sites at high-resolution by AFM. Preliminary results of mapping plasmids (3-7 kb), and lambda DNA (50 kb) by this direct imaging technique, suggest that this technology could be applied to map cosmid-sized clones at accelerated rates.

The potential for E*co*RI restriction mapping larger clones, such as BAC's and YAC's is frustrated only by the molecules becoming entangled on the mounting surface. However, by integrating techniques involving anchoring one end of a DNA molecule to a surface and straightening the molecule by either shear or electrophoretic forces, these larger DNA's could also be mapped by direct AFM imaging. As a consequence existing gaps in contig maps, due to DNA fragments that resist being cloned, would be eliminated.

Improvements in commercial AFM instrumentation combined with our efforts to streamline sample preparation have made this a rapid and reliable technique. This coupled with improvements making existing software more "user friendly," by integrating image acquisition with image analysis, should make this new mapping technology readily transferable to laboratories lacking experience with scanning probe microscopes. Using preliminary results as a benchmark, this new methodology should be capable of mapping *one cosmid per day.*

(1) D. J. Wright, K. King, and P. Modrich, "The negative charge of Glu-111 is required to activate the cleavage center of E*co*RI endonuclease," *J. Bio. Chem.* 264(20), 11816-21, 1989.

---

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

APPLICATIONS OF QUANTITATIVE DNA FIBER MAPPING (QDFM) IN PHYSICAL MAPPING AND SEQUENCING*.

Heinz-Ulrich G. Weier, Mei Wang, Jan-Fang Cheng, Yiwen Zhu, Herbert W. Moise, Christopher H. Martin, Micheal J. Palazzolo and Joe W. Gray, Resource for Molecular Cytogenetics and Human Genome Center, Life Sciences Division, 74-157, University of California, Lawrence Berkeley National Laboratory, Berkeley, CA 94720.

'Quantitative DNA Fiber Mapping' (QDFM) recently developed in our laboratory combines three techniques (MOLECULAR COMBING, FISH and IMAGE ANALYSIS) for precision mapping of individual DNA fragments and determination of distance or overlap between pairs of cloned DNA fragments.

In MOLECULAR COMBING, a solution of purified DNA molecules is placed on a silanated glass slide prepared so that the DNA molecules slowly attach at one or both ends. The DNA solution is then spread over a larger area by placing a coverslip on top, DNA molecules are allowed to bind to the surface and dried. Individual DNA molecules are straightened and uniformly stretched during drying by the hydrodynamic action of the receding meniscus. The position of specific sequences along the stretched DNA molecules is visualized by overnight fluorescence in situ hybridization (FISH) and measured by digital IMAGE ANALYSIS techniques on images recorded from the fluorescence microscope.

In pilot experiments, we applied QDFM to map γδ transposons, plasmid or cosmid probes along P1 molecules, and P1 or PAC DNA clones along straightened YAC molecules ranging in size from ~490kb to >1Mbp[1]. Our studies demonstrated the power of QDFM by showing that

(1) molecular combing and high resolution physical mapping can be performed on DNA molecules linearized by digestion with restriction enzymes, randomly broken DNA or circular molecules,
(2) linear DNA molecules ranging in size range from 17kb to >1Mbp are uniformly stretched to ~2.3kb/μm so that measurements obtained by image analysis [measured in μm] can be converted directly to genomic distances [measured in kb],
(3) only few (<10) molecules are needed for analysis,
(4) the hybridization efficiency is high so that DNA fragments of less than 1kb can be mapped,
(5) plasmids (such as ~3kb sequencing templates), mobilized transposons and cosmid probes can be mapped to within ~2-5 kb along P1 molecules of ~55-95 kb,
(6) the extent and orientation of overlap between two P1 DNA molecules can be determined to within ~3 kb by hybridizing DNA from one clone onto linearized molecules of the other,
(7) the map position of two independent P1 clones along a YAC molecule can be measured with a resolution of a few kb,
(8) the physical distance between two P1 clones representing the proximal ends of two contigs (gap region) can be measured with kilobase resolution, and
(9) the chimerism status of sequencing templates can be determined rapidly.

The impact of QDFM on genome research will depend on how well it scales-up to accommodate the needs of large-scale mapping and sequencing projects. Preliminary results showed that as many as 20 clones can be combed on a single microscope slide. Furthermore, QDFM is highly amenable to automation, which might increase its throughput by orders of magnitude. Molecular combing and FISH require only little user interaction and instrumentation for slide handling (washes, staining etc.) exists. Development of semi-automated slide scanning and image acquisition/analysis should facilitate these aspects of the analysis procedure. These developments should bring QDFM to the point where it is of major utility in assembly of sequence-ready physical maps and quality control during the sequencing process.

*Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research, Department of Energy, under contract DE-AC-03-76SF00098.

[1] H.-U.G. Weier et al. *Human Molecular Genetics* **4**, 1903-1910 (1995).

ANALYSIS OF METAPHASE CHROMOSOMES

Damir Sudar[1,4], Steve Lockett[1], Mark de Kanter[3], Kasper Ligtenberg[3], Gus van der Feltz[3], Dan Pinkel[2,1], Joe Gray[2,1]

[1]Resource for Molecular Cytogenetics; MS 74-157; Lawrence Berkeley National Laboratory; 1 Cyclotron Road; Berkeley, CA 94720. [2]University of California, San Francisco, CA. [3]Delft University of Technology, Delft, The Netherlands. [4]Corresponding author

A number of important molecular cytogenetic analyses depend on the analysis of images of metaphase chromosomes. This includes applications such as DNA probe mapping relative to chromosome bands or as a fractional location along the length of the chromosome [1], CGH chromosome ratio profile measurements [2], translocation detection, and conventional banding-based karyotyping. These analyses rely on the accurate segmentation of chromosomes from the background in the image, decomposition of clusters of chromosomes, and proper determination of chromosome borders. Locations of features such as bands in karyotyping, DNA probes in probe mapping, increases and decreases in CGH analysis, and breakpoints in translocation analysis need to be calculated with high accuracy which is complicated by chromosome bending and differential contraction. We have developed algorithms specifically suited for the segmentation of metaphase chromosomes in fluorescently labeled images and for the accurate determination of location of features along their length.

Automatic segmentation is performed in two stages: presegmentation of the metaphase image, and detection and decomposition of the clusters in the image. In the presegmentation, two parallel strategies are employed: one using operations from (binary) mathematical morphology, the other using a digital Laplace filter for edge detection. These two approaches are combined to get a segmentation along the most likely edges. Clusters of chromosomes are detected based on morphology and size. They are decomposed using on a rule-based decision system between likely cut-points on the contour of the cluster. Likely cut-points are determined from the contour curvature and skeleton branching. Segmentation results of 94% of the chromosomes in a metaphase were achieved for DAPI stained human metaphase spreads.

Mapping of locations along the length of individual segmented chromosomes was achieved by calculating the skeleton of the segmentation mask, converting the skeleton to a smoothed piece-wise linear line description (medial axis), extension to the telomeres of the chromosome, and the extraction of an integrated profile over the width at unit-step locations along the medial axis.

In order to correct for errors in the telomere location, differential stretching of the chromosomes, and differences in chromosome lengths we implemented a recursive warping algorithm of the profiles which converts a profile to a template based on 'matchable' locations along the profiles such as centromeres, telomeres, and chromosome bands. This allows the combination of profiles from multiple metaphase by averaging which is essential for analysis methods such as CGH.

This work was funded by the US DOE contract DEAC0376SF00098.

[1] Mascio LN, Verbeek PW, Sudar D, Kuo W-L, Gray JW, Semiautomated DNA Probe Mapping Using Digital Imaging Microscopy: I. System Development. Cytometry 19:51-59, 1995

[2] Piper J, Rutovitz D, Sudar D, Kallioniemi A, Kallioniemi O-P, Waldman F, Gray JW, Pinkel D. Computer Image Analysis of Comparative Genomic Hybridization. Cytometry 19:10-26, 1995

Image Analysis of Thick, FISH-Labeled Solid Tumor Specimens

Stephen Lockett[1], Anton Rutten[2], Damir Sudar[1], Ramin Khavari[3], Daniel Pinkel[1], Joe Gray[1]

[1]Resource for Molecular Cytogenetics, MS 74-157, Lawrence Berkeley National Laboratory, Berkeley, CA 94720. [2]Collaborating author, Delft University of Technology, The Netherlands. [3]Collaborating author, The University of California, San Francisco.

Fluorescence in situ hybridization (FISH) is a technique for labeling specific nucleic acid sequences inside cells and has wide spread use for analyzing the genetics of tumors. When FISH is applied to solid tumors, it is important to preserve the cellular organization of the tissue so that genetic abnormalities in one cell can be compared to neighboring cells. The standard approach for undertaking such studies is to prepare thin (~ 4 μm) tissue sections, perform FISH and observe the labeled sequences with an epi-fluorescence microscope. However, this approach has the drawback that nearly all cells and nuclei have been truncated by the sectioning process and so it is impossible to accurately determine the number of copies of the sequences in each cell. The aim of this project is to develop FISH protocols, 3D (confocal) microscopy and 3D image analysis for studying 20 μm thick sections where most of the cells are intact.

At this meeting last year we presented interactive algorithms for rapid, visual-based enumeration of punctate FISH signals in each intact cell (Lockett et al). Since then, we have tested these algorithms, begun automation of the enumeration procedures and developed a technique for correcting images for autofluorescence (Szöllösi et al).

The interactive algorithms were tested using control specimens of normal skin where two FISH signals per intact nucleus were expected. We used FISH probes to specific chromosome centromeres and to the 20q13 region (a region commonly amplified in breast cancer) and detected over 80% of the targeted sequences in the intact nuclei.

We began automation of the enumeration procedure by developing an algorithm for detecting fluorescence stained nuclei from 3D images. The algorithm first calculated threshold intensities that were used for dividing the images into bright regions representing nuclei and the dark background. Next, the size and shape of each nuclear region was measured. Regions that were too large or irregular in shape to represent individual nuclei were split into smaller regions using erosion operations. The algorithm detected 75% of nuclei in thick, normal skin of which over 99% were deemed correctly segmented based on visual comparison with the acquired images.

Autofluorescence severely limits sensitivity when detecting FISH signals in tissue sections. An algorithm was written which calculated the autofluorescence component of images of FISH signals using images of only the autofluorescence in the same microscopic scene. After subtracting the autofluorescence component from the FISH images, signals barely detectable before correction were clear visible after correction. Use of the algorithm proved essential for the detection of locus-specific probes (erbB2 and 20q13) in tissue sections.

In this project, we developed (and will continue to develop) technology that for the first time can analyze the genetics of small, premalignant lesions, and directly correlate this molecular cytogenetic information with tissue histology. A major application of this technology will be to follow the evolution of genetic instability during the progression of breast cancer by enumerating nucleic acid sequences in histologically defined regions of tumors that presumably represent different stages of the disease.

This work was funded by the US DOE contract DEAC0376SF00098 and a grant from the Whitaker Foundation.

Lockett, S.J., Thompson, C., Sudar, D., Mullikin, J., Hyun, B., Kharvari, R., Pinkel, D. and Gray, J. (1995) Interactive Algorithms for Rapid Chromosome Copy Number Enumeration of Individual Whole Cell Nuclei Inside Intact Tissue Specimens. Proc. SPIE.

Szöllösi, J., Lockett, S.J., Balázs, M. and Waldman, F.M. 1995. Autofluorescence Correction for Fluorescence in Situ Hybridization. Cytometry. Accepted

ANALYSES OF DUPLICATIONS IN THE HUMAN GENOME BY USING FLUORESCENCE IN SITU HYBRIDIZATION*

Barbara J. Trask, Hillary Massa, Cynthia Friedman, Lee Rowen, Hiroki Yokota, Ger van den Engh, Heather Christy, Leah LaTray, Shawn Iadonato, David Wong, Forrester Johnson, Carolyn Akinbami, John Blankenship, Eric Green[2], Mark Keating[3], Antonia Martin-Gallardo[4], and David Miller[5], Dept. of Molecular Biotechnology, University of Washington, [2]National Center for Human Genome Research, NIH, [3]Dept. Human Genetics, University of Utah, [4]Centro Nacional de Biotecnologia, Madrid; [5]Dept. of Electrical Engineering, Pennsylvania State University.

Low-copy repeats can present a challenge to mapping efforts. This challenge is severe when the repeated segments are large (>kbp), very similar, or closely spaced in the genome. This challenge must be met because these sequence duplications can have biological significance as members of gene families, mediators of chromosome rearrangements, indicators of steps in chromosomal evolution, or descriptors of human variation. Here we present analyses of three biologically interesting regions whose maps have been complicated by low-copy repeats.

a) A collaborative effort (BT, EG, and MK laboratories) is underway to physically map the Williams region of 7q11.2, to correlate patient deletions at the molecular level with the phenotypic features, and to identify and characterize the genes contained within the deleted intervals. These studies have demonstrated that the region is associated with duplications of sequences within a single cytogenetic band. Some parts of the region are also duplicated on other chromosomes. The results of a combination of FISH analyses of patient and normal chromosomes and STS-content mapping of YACs will be presented.

b) We have used FISH to map the boundaries of the duplication of part of the T-cell receptor locus (chromosome 7) on chromosome 9. This duplication encompasses a trypsinogen gene.

c) Different challenges are presented by the low-copy duplications identified by cosmid 7501. The 35-kbp segment cloned from chromosome 19 is inserted at the ends of chromosomes 3, 15, and 19 in diverse human populations. Fifteen other chromosomes carry this sequence in at least one of 45 individuals analyzed, but the sequence is restricted to one chromosome in chimps and gorillas. The duplicated segment contains three regions with high homology to olfactory receptor genes. The extent of the duplication and the sequence similarity of the different copies has been analyzed by a combination of physical mapping approaches (of YACs, P1s, cosmids), molecular analyses of different chromosomes purified by flow-sorting, and FISH. Preliminary data suggest that parts of the cosmid may be duplicated within short YAC-sized segments of the same chromosome. These analyses may reveal the rearrangements chromosomes have undergone during evolution and a more complete description of human variation.

d) FISH analyses of straightened DNA fibers offers the possibility of detecting and analyzing very closely spaced duplications. We will report on some of the parameters that are critical to obtaining uniformly and reproducibly straightened DNA fibers for FISH.

*Supported by a grant from the Director, Office of Energy Research, OHER, U. S. Department of Energy under contract DE-FG06-93ER61553 and DE-FG06-93ER61662.

Technologies for Automated Genome Mapping and Sequencing*

Chip Asbury, David Basiji, Kelly Dillon, Rich Esposito, Curran Fey, Rene Gelderman, Steve Knowles, David Makihara, Dennis Siemer[1], Todd Smith, Barb Trask, Ger van den Engh
Department of Molecular Biotechnology, University of Washington,Seattle, WA, and [1]V-TEK, Mankato, MN.

We are developing automated instruments for several stages of the mapping and sequencing process.

One project concerns the sub-cloning of a BAC or cosmid into M13. The aim is an automated process that will select and expand M13 sub-clones from a library. The daily capacity of the instrument will be sufficient to completely sequence a BAC or cosmid. Individual transformed bacteria are selected by fluorescence-activated cell sorting. Bacteria are deposited into individual culture wells. The turbidity of the wells is continuously monitored providing quality criteria for the growth conditions. The bacterial cultures are automatically harvested when the cultures reach an optimum density within a preset time window. Cultures that do not meet the growth criteria are discarded.

We are also developing an electrophoresis-based extraction technique that will yield DNA of sufficient quality for a sequencing reaction. Two different isolation methods are under investigation. One of them uses dipole induction to separate the DNA from other organic components. The DNA extraction machine connects to the clone isolation and expansion machine. The samples that come out of the extraction process are stored in a sealed plastic ribbon.

We have developed a gel loader that automatically injects DNA samples into an electrophoresis is gel. The prototype reproducibly loads samples onto an agarose gel with a density of one lane per 2 mm. A version that is capable of loading samples onto thin polyacrylamide gels is under construction.

A gel scanner for evaluating restriction fragment gels has been developed. The scanner evaluates that gels that have been prepared in the automatic sample loader. Present activities concentrate on reducing the background signal from the gel.

In all these instruments, the samples are transported in wells in a sealed plastic ribbon. This system will form the basis for a DNA sample repository that can store and access a large number of samples in a small space. The advantages of storing DNA and bacteria samples in plastic ribbons will be discussed.

*Supported by a grant from the Director, OER/OHER of the U. S. Department of Energy under contract DE-FG06-93ER61662 and DE-FG06-93ER61553.

SEQUENCE-READY PHYSICAL MAPS GENERATED BY MULTIPLE-COMPLETE DIGEST MAPPING*

Jun Yu, Gane Ka-Shu Wong, Edward C. Thayer, and Maynard Olson, Department of Molecular Biotechnology, University of Washington, Seattle WA 98195

We have reduced to practice a standardized method for generating unusually accurate, high-resolution restriction maps supported by densely overlapping cloned coverage of the target region. In the present implementation, the starting point is a yeast-artificial chromosome (YAC) contig that has been mapped by STS-content mapping. A subset of the YACs is subcloned at high redundancy into cosmids. The method is equally compatible with other high-level mapping strategies and cloning systems. The main prerequisite is a source of densely overlapping clones from the target region whose DNA can be characterized by cleavage with a variety of restriction enzymes that produce a moderate number (i.e., approx. 5-15) of fragments.

Two principal software systems underlie multiple-complete-digest (MCD) mapping. Will Gillett (Department of Computer Science, Washington University) has developed a sophisticated map-assembly tool, which converts fragment-size lists for a set of densely overlapping clones, each digested independently with an arbitrary number of enzymes, into MCD maps.[1] To provide input for the Gillett map assembler, we have developed a fully automated image-analysis tool that takes digitized images of stained agarose gels on which restriction digests have been fractionated and produces fragment-size lists. Without interactive intervention, this cross-platform tool presently interprets >95% of gel lanes correctly (i.e., no false-positive or false-negative fragment calls) with a typical fragment-sizing accuracy of ±1-2%.

We have reduced this system to practice in a collaboration with Daniel Geraghty (Fred Hutchinson Cancer Research Center), starting with a YAC contig that he has produced for the class I region of the HLA locus.[2] A subset of the YACs is subcloned into cosmids at 20-fold redundancy. Cosmids containing human DNA are selected by colony hybridization using labeled human DNA as the probe. Each cosmid is independently digested with 3 different restriction enzymes that have 6-bp recognition sites. The restriction fragments are separated by size on an agarose gel, which is post-stained (SYBR Green from Molecular Probes), imaged with a fluorescence scanner (Molecular Dynamics FluorImager 575), and analyzed with the software package described above. Vector-insert fusion fragments are identified by gel-transfer hybridization. All insert fragments larger than 500 bp are used by the Gillett mapping software to construct maps which position both restriction sites and clone ends. Initial mapping experience now includes several HLA YACs. In some cases, continuous sequence data are now available from D. Geraghty for mapped regions longer than 100 kbp. In these cases, there is near-perfect agreement between the MCD- and sequence-derived maps. Hence, MCD mapping provides a powerful check on shotgun-sequence assemblies, and allows precise localization of any residual gaps in assembled sequence.

*Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy under contract DE-FG06-92ER61487.

[1]W. Gillett, L. Hanks, G. K.-S. Wong, J. Yu, R. Lim, and M.V. Olson, *Genomics*, in press (1996).

[2]D. Geraghty, *Curr. Opin. Immunol.* **5**, 3-7 (1993).

MICROGENETICS: A NEW METHOD FOR SCREENING LARGE INSERT LIBRARIES*

Saika Aytay,[1] Maureen Thornton,[2] Lisa Davis,[1] Charles Helmstetter,[2] and John Hozier[1]
[1]Applied Genetics Laboratories, Melbourne Florida, [2]Department of Biology, Florida Institute of Technology, Melbourne, Florida 32901.

MICROGENETICS™ is a new method for screening large insert libraries with current emphasis on screening yeast artificial chromosome (YAC) libraries. The combination of techniques under development allows library screening at the microscopic level. We take advantage of a well-established culturing technique, the "baby machine", in which a population of dividing "mother" cells is attached firmly to a surface. In MICROGENETICS the progeny cells are collected as microclones on a parallel surface in a pattern identical to the original, thus forming a replica. Cells on the replica surface are fixed in-situ and sequence-specific probes are hybridized to corresponding YAC DNAs using standard FISH techniques. Alternatively clones of interest are identified by in-situ PCR. Positive microclones are detected by fluorescence microscopy and the viable positive mother cell is removed from the mother plate with a micromanipulator-driven micropipette, and cultured and propagated for further analysis. With MICROGENETICS a YAC library of 30,000 clones can be maintained and screened on a single microscope slide providing a many-fold increase in efficiency over current technology.

During the process of developing this technology we have used several procedures to attach the yeast cells to glass slides, since normally yeast cells do not attach to glass. We have coated slides with poly-lysine, Cell-Tak, silane and Con-A. Individual colonies remain attached to Con-A treated slides during the process of in-situ PCR and fluorescence detection if the cells are fixed with methanol and acetic acid after replica formation. We have performed in-situ PCR on replica colonies using several sets of primers with unique sequences which target specific dinucleotide repeat markers in known mouse YAC clones which were characterized in earlier studies. With in-situ PCR we obtained the same patterns of positive and negative results with respect to markers used for each YAC clone as predicted by our previous studies. We have also spiked YAC clones known to be negative for a given primer set with positive YAC clones in known ratios and proceeded with in-situ PCR and fluorescence detection, and observed the expected ratio of positive to negative signals. Similar experiments are being performed to detect positive colonies at lower frequencies which will mimic positive colonies in standard YAC libraries. We are also forming gridded arrays of known positives in a background of negatives. Lastly, we will isolate mother cell microcolonies for propagation and show that they are indeed the positive colonies which we spiked among negative colonies.

MICROGENETICS technology can be applied readily to other large insert libraries with bacterial hosts (BAC's, P1's), since the "baby machine" technique was originally designed for E.coli. This technology reduces a given library from a stack of hundreds of microtiter plates to a single surface and gives a clone individualized treatment only after identification of the clone of interest in the screening step.

*Supported by grant DE-FG-02-95ER 81924 from the Director, Office of Health and Environmental Research, Health Effects and Life Sciences Research Division of the U.S. Department of Energy.

INTERACTIVE SEMI-AUTOMATIC ASSIGNMENT OF MULTIPLE PROBES TO CYTOGENETIC BANDS BY SIMULTANEOUS DUAL COLOR FLUORESCENCE IN SITU HYBRIDIZATION AND DAPI BANDING*

S. Burde, G. Joss[1], *J. A. Gonzales, C. H. Coulon, L. L. Deaven and B. L. Marrone,* Los Alamos National Laboratory, Life Sciences Division, LS-5 M888, Los Alamos, NM 87545

A macro was developed to run in conjunction with the popular image analysis package NIH Image to allow simultaneous determination of mapping positions of one or two separate DNA probes with respect to cytogenetic bands by dual color fluorescence in situ hybridization (FISH) and DAPI(4,6-diamidino-2-phenyl-indole dihydrochloride) banding.

Chromosomes were hybridized with cosmid probes labeled with fluorescein-11-dUTP or biotin-12-dUTP detected with Streptavidin-Texas Red. Chromosomes were counterstained with DAPI, and separate images of DAPI-fluorescence and each probe fluorescence were acquired using a Zeiss Axiophot microscope fitted with a Photometrics slow-scan cooled CCD camera containing a 1 K x 1 K Kodak KAF-1400 Grade 1 chip. Images were subsequently converted into rgb-color tiff stacks and contrast-stretched for analysis.

In order to allow maximal flexibility, a user-defined line along the chromosome is used for measurements. Algorithms were developed to detect the ends of the chromosome and the cytogenetic bands. Results of the analysis are presented in graphical form, comprising a display of the DAPI intensity along the chromosome, the positions of the probe(s), the locations of bands as determined by analysis of the second derivative of the DAPI intensity profile, and a standard ideogram of the chromosome for comparison.

The approach was validated and compared to visual assignment of probes to DAPI bands using the cosmid clone PYGM which has been previously mapped to chromosome 11q13 [2]

*Supported by US DOE (W-7405-ENG-36)

[1]School of Biological Sciences, Macquarie University, North Ryde (Sydney), NSW 2109 Australia

[2] C. Junien, V. van Heyningen, G. Evans, P. Little and M. Mannens: Report of the Second Chromosome 11 Workshop: Genomics 12: 620-25 (1992)

Sizing DNA Fragments by Flow Cytometry and Applications to the Analysis of Cloning Vector Inserts

J. T. Petty[†], Z. Haung[†], R. Habbersett[††], J. H. Jett[††], and R. A. Keller[†]

[††]Chemical Science and Technology Division and [†]Life Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

We have demonstrated that our flow cytometry based ultrasensitive fluorescence detection can be used to size DNA fragments ranging from 564 base pairs to 167 kilobase pairs. Size analysis is accomplished by staining the fragments with intercalating dimeric nucleic acid stains, POPO3 or TOTO, which bind stoichiometrically to DNA. When the stained fragments pass through the exciting laser beam, the amount of fluorescence emitted is proportional to the number of base pairs in the fragment.

Small size range samples consist of restriction digests of Lambda phage DNA along with the complete Lambda genome. Large fragment samples consist of a Lambda digest, full length Lambda, and the T4 and T5 phage genomes to provide fragments ranging in size from 17 to 167 kilobases. Samples are analyzed at a concentration (10^{-13} M) at which the probability of more than one fragment being present in the probe volume at any one time is very small. Data are collected by recording the time history of detected photons. The integrated intensities from the individual fragments are assembled in a histogram. Histograms are fitted to a sum of Gaussians plus an exponential background function and the centroids of the peaks are plotted versus the known fragment sizes. Resulting calibration curves are linear over the size ranges analyzed. This result is to be compared with the pulsed field gel electrophoresis which results in a highly nonlinear calibration curve.

The size range of DNA fragments that have been analyzed by these techniques is from 0.564 kilobases to 167 kilobases, covering over 2 orders of magnitude. Data collection is accomplished in less than three minutes. Since each fragment is counted, an absolute measure of the number of fragments in each size class is obtained from the measurement. The amount of material required to produce a statistically well defined distribution is small. Data are collected for approximately 10,000 fragments which translates to a mass of approximately 2×10^{-13} grams for the small fragment size range.

We are currently constructing a simple, inexpensive apparatus to perform these measurements. Excitation will be accomplished with a 30 mw diode pumped Nd:YAG laser that emits at 532 nm. Photon detection will be with a solid state avalanche photodiode which provides a logic signal to a multichannel scaler board that resides in a PC. The whole apparatus will be contained in a cubic foot plus the personal computer.

Applications of fragment sizing to the analysis of the size and stability of vector inserts will be discussed.

This work supported by Los Alamos National Laboratory LDRD funds and the DOE/OHER Human Genome Program.

CELL BY CELL FLOW ANALYSIS OF HUMAN CHROMOSOME SETS

V.V. Zenin, N.D. Aksenov*, A.N. Shatrova*, N.V. Klopov#, L.S. Cram^, A.I. Poletaev °; *Institute of Cytology , Russian Academy of Sciences, St.-Petersburg, Russia; St.-Petersburg Institute of Nuclear Physics, Russian Academy of Sciences, Gatchina, Russia; ^Los Alamos National Laboratory, Los Alamos, NM; °Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia*

Instrumentation for univariate fluorescent flow analysis of chromosome sets has been developed for human cells. A new method for cell preparation and intracellular staining of chromosomes with different dyes was developed. The method includes enzyme treatment (chymotrypsin), incubation with saponin and separation of prestained cells from debris on a sucrose gradient. This procedure makes it possible to get a well stained sample with a minimal amount of contaminates: free chromosomes, cell fragments, etc.
A special mixing/stirring device was placed inside the flow chamber of flow cytometer. The rupturing of prestained mitotic cells is performed by means of a small magnetic rod vibrating in an alternating magnetic field. The device works in a stepwise manner: a defined volume of sample is delivered to the breaking chamber for breaking mitotic cells for a defined time period, followed by a buffer wash to move the released chromosomes from the breaking chamber to the point of analysis. The information about the chromosomes appearing at the point of analysis is accumulated in list mode files makes it possible to resolve chromosome sets arising from single cells on the basis of time gating. The concentration of cells in the sample must be kept low to ensure that only one cell at a time enters the breaking device. The developed software classifies chromosome sets according to different criteria: total number of chromosomes, overall DNA content in the set, and the number of chromosomes of certain type. In addition it's possible to determine the presence of extra chromosomes or the loss of chromosome types. Thus this new approach combines the high performance of flow cytometry (quantitation and high throughput) with the advantages of image analysis (cell to cell karyotype analysis and the skills of a trained cytogeneticist). The data analysis capabilities offer extensive flexibility in determining important features of the karyotypes under study. This development offers the potential to duplicate most of what is determined by a clinical cytogeneticist.

Supported by a grant from the DOE-NIS International Partnering Program, Russian State Program of Human Genome Studies, and the NIH (grant RR-01315).

Chemiluminescent Detection of Multiplex Labeled Microsatellite Markers and DNA Sequences and Evaluation of New Membrane Surfaces

Chris S. Martin, John C. Voyta and Irena Bronstein, Tropix, Inc., Bedford, MA 01730

We have developed a technique for sequential nonisotopic detection of multiple sets of DNA reaction products which are labeled with different haptens. This multiplex labeling approach utilizes hapten-specific alkaline phosphatase conjugates and chemiluminescent 1,2-dioxetane substrates. For DNA sequencing, multiple primers, each with a unique ligand label, are incorporated in sequencing reactions, the products are separated, transferred to nylon membrane and detected by binding hapten specific alkaline phosphatase antibody conjugates. We have used primers labeled with biotin, digoxigenin, fluorescein and 2,4-dinitrophenyl (DNP), enabling the acquisition of four images of DNA sequence data from a single nylon membrane. The need for large scale screening of polymorphic microsatellite markers for genetic mapping led us to adapt this technique for the detection of PCR amplified microsatellite markers. Individual sets of PCR primers labeled with each hapten are utilized to amplify different microsatellite repeat markers. The amplified markers for each genomic DNA sample are loaded in a single gel lane, electrophoretically separated, transferred to a nylon membrane and detected sequentially with hapten-specific alkaline phosphatase conjugates. Each of the four different haptens have been used for three amplimer pairs, to generate three different size fragments with each label. Thus, 12 different markers can be typed from a single gel lane. While satisfactory chemiluminescent images can be obtained on nylon membranes, there is a need for improvement of the membrane support used for chemiluminescent detections. Recently, we developed membranes which incorporate a polymer enhancer surface layer. These membranes, prepared by overcoating certain membrane supports, exhibit greater chemiluminescence signal intensities and lower background noise. These membranes were tested by performing chemiluminescent detection of multiple labeled oligonucleotides or DNA sequence ladders. The development of a superior membrane would enable more rapid imaging of DNA sequences on x-ray film or electronic imaging devices (i.e. CCD cameras).

This work was funded by the DOE Genome Program.
Contract No. DE FG05 92ER81389

DETECTION OF INELASTICALLY-SCATTERED LIGHT WITHIN A FREE STREAM IN AIR IN FLOW CYTOMETRY

Raymond Mariella Jr., Richard Langlois, Donald Masquelier, Mukund Venkatesh, Shadi Shakeri, Gerald Eveleth, and Dino Ciarlo. Human Genome Center, Biology and Biotechnology Research Program, L-452, Lawrence Livermore National Laboratory, Livermore, CA 94550

Girish Vyas, K.S. Venkateswaran, Univ. Cal. San Francisco, Dept. of Laboratory Medicine, Parnasus St., San Francisco, CA 94143

We have recently invented[1] a new technique to collect perpendicular light scatter, both elastic and inelastic, by using the total-internal reflection and optical waveguide properties of a free stream in air (a "flow-stream waveguide"). This produces a 0.88-numerical aperture optical collection and very high optical transmission efficiency to the optical detector. It is easier to align than systems which use microscope objectives and it is less expensive, as well. We extract the light from the aqueous flow stream by placing a conically-polished fiber optic in the stream with the other end of the fiber optic facing an optical detector, with or without wavelength-dispersing optics between the fiber and the detector. We will be reporting our studies of a variety of optical configurations for low-light-level detection, including the use of a small monochromator for spectral analysis. With the small monochromator, we have observed the level of fluorescence of the 3M, Inc. TECS® fiber which is generated by pulses of elastically-scattered light. This, of course, is an undesirable effect. We have also used this system to measure the emission spectrum of fluoresceinated latex beads as part of a collaboration between LLNL and UCSF. for both this collaboration and for internal use for the Genome Project we are examining the limiting features (such as fluorescence of the fiber optics!) with the goal of achieving the detection of weakly fluorescent particles, such as DNA which has undergone digestion by a restriction enzyme.

One unexpected and very pleasant discovery as part of our collaboration with UCSF researchers is a technique to characterize unfixed erythrocytes for A and B surface antigens as well as Rh factor, all performed in a single tube and determined rapidly by flow cytometry.

We have also used the flow-stream-waveguide technique to evaluate a sheath-flow/sample injector which has been fabricated from precision-etched silicon wafers which have been bonded together to create the flow channels. Although it did work the first time we tried it, this microfabricated sheath-flow/sample injector was significantly more difficult to operate than a traditional commercial version.

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

[1] Patent application claims allowed by US Patent Office, June 1995.

[2]Shadi Shakeri, K.S. Venkateswaran, R. Mariella Jr., G.N. Vyas, Human ABO and RH Blood Typing by Simultaneous Three-Color Cytofluorometry, to be presented at the Int. Soc. Blood Transfusion, April 1996, Tokyo

The Advantages of a Miniature PCR Instrument Based on Micromachining Technology

M. Allen Northrup, Stacy Lehew, Phoebe Landre, Peter Krulevitch, Bart Beeman, Dean Hadley, and Bill Bennet, Microtechnology Center, L-222, Lawrence Livermore National Laboratory, Livermore, CA 94551

Significant advantages can be attained by miniaturizing components of diagnostic instruments. Theory predicts huge gains can be made in efficiency and speed of analysis for chemical separation systems such as those used in chromatography and electrophoresis, and several research groups are taking advantage of these favorable scaling laws. Similar advantages are afforded by the miniaturization of chemical reactors allowing for new levels of performance and efficacy. We will show how these advantages are being used to build a miniature, low-cost, low-power, and high efficiency PCR instrument.

In this report we detail the design and development of a hand-held, low-power, feedback-controlled thermal cycling instrument for performing the polymerase chain reaction (PCR) that uses microfabricated, silicon-based reaction chambers. Several different reaction chamber designs have been modeled, built, and tested. Each design incorporates an integrated thin film heater, passive silicon cooling surfaces, and optical windows for detection of the reaction. A highly efficient, battery-operated controller has been implemented that shows significant improvements over commercial thermal cycling instrumentation. Several different biological systems have been detected with the miniature PCR instrument including viral, bacterial and human genomic DNA targets. We have also performed amplification of human DNA targets that are specific for the disease, cystic fibrosis. This is a multiplex amplification system (i.e. amplifies 8 different sections of DNA simultaneously) and requires extremely precise temperature control. We have been able to provide the requisite control with the miniature system for the eight CF mutations on human DNA which were subsequently verified on simple test strips. The significance of these results are that for the first time a hand-held, battery-operated instrument along with simple test strips can be used to detect an important genetic disease. Recent results include low power, ultra-fast thermal cycling where we have been able to amplify DNA targets in 30 cycles in less than 7 minutes. We have also been able to monitor the reaction real-time in the miniature instrument using fluorescence monitoring with a miniature, low power optical system based on diodes. This system has also been able to detect less than 30 picomolar concentrations of fluorescently-label DNA primers. These results indicate a new ability to perform detailed studies of the reaction kinetics and improve the efficiency of this important diagnostic technique. Due to the use of microelectromechanical systems (MEMS) technology, we have shown that low- cost, high efficiency, biotechnological and clinical diagnostic instrumentation is a reality.

(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.)*The authors acknowledge the support of the MEMS program of the Advanced Research Projects Agency. We would also like to acknowledge the collaboration of Roche Molecular Systems of Alameda, Ca.*

This page intentionally left blank.

Informatics

THE GENOME DATA BASE V6*

Kenneth H. Fasman, Stanley I. Letovsky, Peter Li, Robert W. Cottingham, and David T. Kingsbury, Division of Biomedical Information Sciences, Johns Hopkins University School of Medicine, Baltimore, MD 21205-2236

GDB has been completely redesigned to address some well-recognized shortcomings, such as:

- Text-based display of map information.
- Difficulties casual users have in finding information of common interest, such as gene descriptions.
- Data acquisition and curation being funneled through HUGO committees and GDB staff.

The focus of the development of GDB 6.0 was the redesign of the database schema and front-end interfaces to address these inadequacies, including:

- An improved data representation of genetic and physical maps allowing for graphical map visualization and querying, including display of multiple aligned maps.
- A new curatorial model allowing direct community editing and curation, including third-party annotation to support a greater diversity of opinion about features of the genome and their location.
- An improved model for gene information including links to databases describing function, structure, products, expression, and associated phenotypes.

The development of GDB 6.0 utilizes several new technologies including an object-oriented data model, object broker, data-driven WWW interface, and graphical interfaces for most popular computer platforms.

With 6.0, GDB staff hope to increasingly promote community participation in the design process regarding both the data model and front-end interfaces, so that we can provide the tools necessary to support the Human Genome Project.

Future enhancements to GDB will include improved map editing, an integrated editing environment, querying and browsing, integration with the GSDB Sequence Annotator and the Mouse Genome Database interfaces, and improved polymorphism and mutation representation.

*Supported by the U.S. Department of Energy (DE-FC02-9ER6130), the U.S. National Institutes of Health, and the Science and Technology Agency of Japan, with additional support from the Medical Research Council of the United Kingdom, the INSERM of France, and the European Union.

DATA ACQUISITION AND CURATION IN THE GENOME DATA BASE V6*

Michael A. Chipperfield, Christopher J. Porter, and C. Conover Talbot, Jr., Division of Biomedical Information Sciences, Johns Hopkins University School of Medicine, Baltimore MD 21205-2236

The release of GDB version 6.0 heralds a significant change in the means by which data enter the database, thereby redefining the role of the Data Acquisition and Curation group. Heretofore, data acquisition and entry formed a major part of the activities of the Data group. The vast majority of GDB data, whether submitted by researchers or acquired from other sources, has been entered into the database by GDB staff.

GDB 6.0 opens editorial access to anyone from the wider genome community who requests it. Researchers can themselves enter and modify their own data, and annotate data submitted by others. The focus of the Data group can now move from acquisition to curation. The group will monitor the incoming data to ensure data integrity and quality within the database.

It is vital that electronic bulk submission tools be available, since 90% of the data submitted to GDB currently arrive in electronic form. Improved bulk submission tools have been developed which better reflect the data in the new GDB schema and which load data through the Object Broker (OB) interface. These tools require the use of a new data submission format. In order to ease the transition to this format, tools will be available to translate GDB 5.x style submissions. This will allow submitters time to convert their local systems to the new format. Alternatively, small submissions can be entered directly through the World Wide Web (WWW) interface.

The new design of GDB expands the use of WWW links to the information stored in other databases. The Data group will explore and develop links to other databases such as GSDB, OMIM and protein databases, as well as links to chromosome- or gene-specific WWW pages.

Although the WWW browsing interface is already known to the community, it will take time for them to become familiar with the editorial interface and to begin routinely entering data. We anticipate that some users will continue in the short term to submit their data to GDB for entry. Moreover, the submission of data to GDB is not considered equivalent to the publication of those data in peer-reviewed journals. For these reasons, the Data group will continue its journal scanning activities, while working with HUGO and journal editors to encourage direct submission to GDB.

*Supported by the U.S. Department of Energy (DE-FC02-9ER6130), the U.S. National Institutes of Health, and the Science and Technology Agency of Japan, with additional support from the Medical Research Council of the United Kingdom, the INSERM of France, and the European Union.

A REVISED SCHEMA FOR THE GENOME DATA BASE V6*

Kenneth H. Fasman, Stanley I. Letovsky, Peter Li, Krishna Palaniappan, Michael A. Chipperfield, Christopher J. Porter, John M. Campbell, Edward W. Kraska, Sue E. Borchardt, and Deborah J. Schneider, Division of Biomedical Information Sciences, Johns Hopkins University School of Medicine, Baltimore MD 21205-2236

GDB 6.0 is a family of interrelated data sets operated by the Genome Data Base project. It consists of HGD (Human Genome Database), the mapping data component; CitDB, which holds literature citations; and the Genome Registry, which contains information on people and organizations in the genome community. We view the separation of this information into multiple databases as a pilot effort toward federating genomic databases across the Internet.

The object class hierarchy for the revised GDB starts with *DBObject*. This class contains basic attributes pertinent to all significant objects, such as owner, release date, and accession number. The remainder of the important objects in HGD are divided among five core classes:

GenomeObjects: Things making up or associated with genomes, such as *GenomeRegions*, *GeneFamilies*, and *GeneProducts*. It includes chromosomes, genes, phenotypic markers, cytogenetic landmarks, STSs, and contigs, among others. This is an enhancement of the concept of *Locus* from previous GDB releases.

MapObjects: Data that describe order and distance relations among regions of the genome, as inferred from mapping experiments. Other classes in this category represent higher-order relationships (i.e., alignments) between maps.

ExperimentObjects: Information about mapping experiments, experimental reagents, and experimental results from observed interactions between reagents.

VariationObjects: Data describing mutations, polymorphisms, population frequencies, etc.

AnnotationObjects: Objects that allow users of the database to comment on other objects in GDB. Literature citations, annotations, and cross-references to other databases may be associated with all user-submitted objects in the database.

The new schema was developed using the Object Protocol Model [1]. Databases designed with OPM are easier to create and understand than an equivalent relational database. This is because the relationships between classes and their attributes, and between pairs of classes, are more explicitly defined. OPM allowed the GDB staff to develop the 6.0 design more quickly than would have been possible using the relational model. More importantly, the new database schema is easier for users of GDB to comprehend, and therefore easier to query.

Detailed documentation on the latest database schema can be obtained from GDB's WWW and anonymous FTP servers. We welcome feedback from the community on this and all aspects of the project's design and operations.

[1] Chen, I.A., and Markowitz, V.M. An overview of the Object-Protocol Model (OPM) and OPM data management tools. http://gizmo.lbl.gov/DM_TOOLS/OPM/doc/OPM_3/Overview.ps

*Supported by the U.S. Department of Energy (DE-FC02-9ER6130), the U.S. National Institutes of Health, and the Science and Technology Agency of Japan, with additional support from the Medical Research Council of the United Kingdom, the INSERM of France, and the European Union.

IMPROVED MAP REPRESENTATION FOR THE GENOME DATA BASE V6*

Stanley I. Letovsky, Kenneth H. Fasman, and Peter Li, Division of Biomedical Information Sciences, Johns Hopkins University School of Medicine, Baltimore MD 21205-2236

Version 6.0 of the Human Genome Data Base implements a greatly improved representation of genomic maps over previous releases of the database. The goals of the new design were to improve the expressiveness of GDB's map model, to represent genomic regions at multiple resolutions, to produce map renderings easily, to provide explicit representations of the experimental data and inferred spatial relationships underlying maps, to provide more sophisticated querying on order and distance relationships, to allow for ongoing community contribution to maps, to represent alignments of maps, and to provide support for larger, denser maps of the genome.

The core of the new map representation is the GDB *Map* object, which represents all genomic maps as sets of *GenomeRegions*, each having both a coordinate position and a specified pair of flanking markers. *GenomeRegions* may be points or intervals as appropriate for the map's level of resolution. The flanking markers provide order information that can be used to determine the precision of the coordinate assignment. Order-only maps can be accommodated by using arbitrary, ordinal coordinates. A typical map is a combination of fully ordered "framework" regions and other markers placed within specified framework intervals.

MapRelations represent the component order and distance relationships from which maps are constructed. These relationships among genome regions come in two types: *TwoPointDistances* and *ThreePointOrders*. Map relationships are either transitively derived from other *MapRelations* or are inferred directly from *ReagentRelations*, observed experimental relationships between *MappingReagents*.

The GDB 6.0 representation of maps is designed for efficient database searching. It supports queries on position, order, and distance. One can find all maps or genome regions that overlap, contain, or are contained in a specified genome region. The region can be specified as a single marker, a marker plus or minus a distance, or as a range defined by a pair of markers. One can also find all maps consistent with a specified marker order or inter-marker distance range.

The new map representation is intended to capture better the whole-chromosome genetic and physical maps which are the Human Genome Project's current focus. At the same time, this model sets the stage for a better integration of map and sequence data, as the emphasis shifts to the production of "sequence-ready" maps.

*Supported by the U.S. Department of Energy (DE-FC02-9ER6130), the U.S. National Institutes of Health, and the Science and Technology Agency of Japan, with additional support from the Medical Research Council of the United Kingdom, the INSERM of France, and the European Union.

AN EXTENSIBLE OBJECT BROKER FOR THE GENOME DATA BASE V6

Peter Li, David Waldo, Stuart V. Pineo, and John M. Campbell, Division of Biomedical Information Sciences, Johns Hopkins University School of Medicine, Baltimore MD 21205-2236

The GDB Extensible Object Broker (OB) attempts to resolve two major problems in today's genomic databases. The first is the existence of incompatible architectures among genomic databases, which makes it difficult for the users to collect relevant biological information from different databases. This incompatibility also makes it challenging for the administrators to synchronize internal and external database objects. The second problem is the challenge of managing software development in a rapidly evolving domain where frequent changes of the schema and database technologies require the recoding of the front-end based on new concepts and tools.

OB addresses these two problems with an extensible framework of front-end and back-end modules. The front-end includes the communication, parser, schema, and session modules. The communication module manages the communication channel. It is based on an event-driven paradigm and is designed for graphical user interfaces. The common transport language (CTL) is processed by a parser module that encodes and decodes information between clients and OB. This module is reentrant, multithread-safe, and designed for high throughput. The extensible schema module allows the client to download and query the schema locally, and also incorporates different data model and domain rules. The session module provides a higher-level application programming interface (API) that simplifies development of client applications.

The back-end includes the server DBMS, query, security, and session modules. The server DBMS module interfaces with the native DBMS, such as Sybase. It provides an abstraction for querying and updating the database. The query module translates OB client query language, based on the Object Protocol Model [1], to DBMS commands. The security module provides user login and access control for database objects. It uses robust algorithms for authentication and exportable algorithms for encryption. The back-end session module handles OB resource management and simplifies the porting OB to other environments.

The creation of OB allows GDB to develop a model for the federation of databases. With its architecture based on an extensible framework of modules, it enables independent software development on both front-ends and back-ends. To simplify third-party software development, GDB is providing OB APIs for common languages and platforms. All of the front-end modules are available in C, C++, and Perl; and will be available in other environments as demand warrants. Finally, the OB modules can be extended easily with other interfaces for the front-end and other DBMSs for the back-end.

[1] Chen, I.A., and Markowitz, V.M. An overview of the Object-Protocol Model (OPM) and OPM data management tools. http://gizmo.lbl.gov/DM_TOOLS/OPM/doc/OPM_3/Overview.ps

*Supported by the U.S. Department of Energy (DE-FC02-9ER6130), the U.S. National Institutes of Health, and the Science and Technology Agency of Japan, with additional support from the Medical Research Council of the United Kingdom, the INSERM of France, and the European Union.

USER INTERFACES FOR QUERYING, EDITING, AND VIEWING THE GENOME DATA BASE V6*

Sue E. Borchardt, Stanley I. Letovsky, Kenneth H. Fasman, Laurie C. Kramer, Karen J. Phipps, Lita A. Kearney, Deborah J. Schneider, and Thomas C. Emmel, Division of Biomedical Information Sciences, Johns Hopkins University School of Medicine, Baltimore MD 21205-2236

Prior to the latest release of GDB, users were able to access the database using the APT interface or a World Wide Web (WWW) client. Both of these applications were hand-coded based on the schema and were therefore difficult to maintain. Both applications were further limited by their inability to display graphical query results such as genomic maps. With the release of GDB version 6.0 come new and better ways for users to contribute and view data. Additionally, recognizing the diversity of our users' needs, we have abandoned the approach of providing all functionality in a single application. We are developing different tools for different tasks.

The means for querying and editing GDB in the 6.0 release is provided via a WWW interface. The query and edit forms are generated automatically from the GDB 6.0 schema using the Genera toolkit [1] modified to use an OPM [2] object-oriented schema. The Genera-created application provides a consistent low-level interface that allows users to move easily between query, update, and insert operations. The results of some queries, such as genomic maps, cannot be usefully represented in a generic form-based display. These genomic maps are displayed graphically using a Web external viewer (helper application) that is integrated with Netscape™ to provide the capability to query further on objects contained in a map. This map viewer was written using a cross-platform development tool that allows us to support many hardware platforms while maintaining a single body of code.

We hope the move to more generic, data-driven applications will allow user interfaces to develop more independently of the database. This independence should allow a wider variety of applications, including higher-level query tools, report generators, and graphical editors. We are investigating the use of new technologies such as Sun's Java which will allow us to respond to our users' needs by prototyping applications more quickly. We seek the community's feedback on this new generation of tools to determine what enhancements and additional programs are needed to serve GDB's users in the future.

[1] Letovsky, S.I., and Berlyn, M.B. Genera: A specification driven Web/database gateway tool. http://gdbdoc.gdb.org/letovsky/wgen.html

[2] Chen, I.A., and Markowitz, V.M. An overview of the Object-Protocol Model (OPM) and OPM data management tools. http://gizmo.lbl.gov/DM_TOOLS/OPM/doc/OPM_3/Overview.ps

*Supported by the U.S. Department of Energy (DE-FC02-9ER6130), the U.S. National Institutes of Health, and the Science and Technology Agency of Japan, with additional support from the Medical Research Council of the United Kingdom, the INSERM of France, and the European Union.

THE GENOME SEQUENCE DATABASE (GSDB): MEETING THE CHALLENGE OF GENOME-SCALE SEQUENCING

*Gifford Keen, Jillian Burton, David Crowley, Emily Dickinson, Ada Espinosa-Lujan, Ed Franks, Carol Harger, Mo Manning, Shelley March, Mia McLeod, John O'Neill, Alicia Power**, Maria Pumilia, David Rider, John Rorlich, Jolene Schwertferger, Linda Smyth, Nina Thayer, Charles Troup, and Chris Fields,* National Center for Genome Resources, 1800 Old Pecos Trail, Santa Fe, NM 87505 USA

The Genome Sequence DataBase (GSDB) is a complete, public relational database of DNA sequences and annotation maintained by the National Center for Genome Resources (NCGR). GSDB provides direct, client-server access to the data for data contributions, community annotation, and SQL queries. A multiplatform graphic user interface, the GSDB Annotator, is freely available. Automatically-updated relational replicates of GSDB are also freely available.

GSDB is designed to meet the requirements for a community sequence database outlined by Waterman et al.[1]. GSDB supports complex, *ad hoc* queries in a standard language, SQL[2]. GSDB represents sequence data produced by any strategy, and supports the contribution of additional sequence data or structural or functional annotation by multiple researchers. GSDB extends the Electronic Data Publishing paradigm[3], in which the database is viewed as a primary publication for data not appearing in the traditional literature, to a model in which the database serves as a multi-user laboratory database for the entire molecular biology community. Multiple sequences from a given region and structural and functional annotations on sequences are viewed, in this model, as independent observations, with authors and unique identifiers. GSDB provides multi-user editing capabilities, with the necessary authorship, data security, integrity-checking, and versioning mechanisms needed to ensure that multiple authors do not overwrite each other's work. A mechanism is also provided for individuals or groups to define their own curated views of the data, which include whatever sequences and features they select. Database users may choose to access the entire database, or only the view maintained by an editorial group that imposes particular standards or selects data relevant to a particular set of interests. GSDB functions, therefore, both as a community laboratory database, and as a collection of multiple, virtual, specialty databases.

Information about GSDB and data input and output tools are available at http://www.ncgr.org.

*Supported by Cooperative Agreement 95ER62062 with the U.S. Department of Energy, Office of Health and Environment Research.

[1]Waterman, M. et al., *J. Computational Biology* 1, 173 - 190 (1994).
[2]U.S. Department of Commerce, *Federal Information Processing Standard Publication 127-2: Database Language SQL.* National Institute of Standards and Technology (1993).
[3]Cinkosky, M. et al., *Science* 252, 1273 - 1277 (1991).

REQUIREMENTS ANALYSIS AND FUNCTIONAL SPECIFICATION FOR A SEQUENCING INFORMATION MANAGEMENT SYSTEM

Chris Fields, Gifford Keen, Shelley March, David Rider, John Rorlich, and Charles Troup, National Center for Genome Resources, 1800 Old Pecos Trail, Santa Fe, NM 87505 USA

High-throughput production sequencing operations require robust information management systems to function efficiently and cost-effectively[1]. Key requirements for such systems include the following:

Process integration: The sequencing process includes steps ranging from clone library preparation through both automated and interactive analysis of the resulting sequence data. A robust data management system must represent all steps in this process in a way that allows queries to access and correlate any data or metadata generated during the process.

User-definable procedures: Procedures for materials preparation, sequencing, and data analysis may change weekly. Use of inappropriate or obsolete procedures can cause data inconsistency and render results unusable. The data management system must provide data representations and user interfaces that allow new procedures to be defined, accessed by those implementing them, and tracked as they are applied.

Quality-control and failure analysis: Lack of appropriate data management often seriously impacts quality control. The data management system needs to support both longitudinal and retrospective data quality analyses spanning the entire sequencing and analysis process. The system must also support explicit failure-mode tracking.

Cost accounting and process optimization: Often it is unclear to a laboratory's managers how much alternative processes cost in materials and staff time. Even the locations of relevant bottlenecks are often hard to identify. The data management system must track resource and personnel use and process success and failure rates and conditions in a way that allows appropriate cost accounting and process optimization.

NCGR is developing a detailed requirements analysis and an implementable functional specification for a Sequencing Information Management System with these capabilities. This system will run on a commercial relational database management platform with multiplatform graphic user interfaces, and will include an applications programming interface capable of supporting both public and commercial data analysis tools.

*Supported by Cooperative Agreement 95ER62062 with the U.S. Department of Energy, Office of Health and Environment Research.

[1]Fields, C. in *Automated Technologies for Genome Characterization*, ed. T. Beugelsdijk, Wiley (in press).

THE ROLE OF INTEGRATED SOFTWARE AND DATABASES IN GENOME SEQUENCE INTERPRETATION AND METABOLIC RECONSTRUCTION*

Terry Gaasterland, Natalia Maltsev, Ross Overbeek, Evgeni Selkov, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439

Through PUMA, MAGPIE, and metabolic reconstruction algorithms, we carry genome interpretation beyond the identification of gene products to a customized view of an organism's functional properties.

MAGPIE is a system designed to reside locally at the site of a genome project and actively carry out analysis of genome sequence data as it is generated.[1] DNA sequences produced in a sequencing project mature through a series of stages that each require different analysis activities. Even after DNA has been assembled into contiguous fragments and eventually into a single genome, it must be regularly reanalyzed. Any new data in public sequence databases may provide clues to the identity of genes. Over a year, for 2 megabases with 4-fold coverage, MAGPIE will request on the order of 100,000 outputs from remote analysis software, manipulate and manage the output, update the current analysis of the sequence data,[3] and monitor the project sequence data for changes that initiate reanalysis.

PUMA is a Web-based system offering integrated access to metabolic pathways, multiple sequence alignments, compounds, sequences, and gene products together with a general overview function. Effective interpretation of genomic sequence requires a functional overview, the ability to embed sequence data within a metabolic framework, alignments that integrate specific genes and corresponding proteins within a broader context, and a phylogenetic perspective. Beyond creating an integrated universe of biological data relevant to sequence interpretion, PUMA aims to support customized functional overviews for a large number of organisms. Over 200 such customized overviews are supported in the current release. These 200 include each well-represented organism in the sequence databases. A PUMA functional overview for an organism is generated by projecting the functions that have been assigned to gene products onto a general functional overview.

There are a number of possible perspectives around which general functional overviews can be constructed. The motivating force behind the PUMA functional overview is to create functional slots to hold the 12,000 alignments (from the collection generated and maintained by Randy Smith and his colleagues) and the 1600 metabolic pathway diagrams (extracted and provided by Evgeni Selkov from the Enzymes and Metabolic Pathways database). Since no single perspective determines an obviously superior functional organization, PUMA implements and graphically presents alternative organizations.

Once the functional overview has been established, it remains to pinpoint the organisms' exact metabolic pathways and establish how they interact.[2] This task, which we call **metabolic reconstruction**, begins by producing a set of *established enzymes* (i.e., enzymes with strong similarities in identified coding regions to existing sequences for which the enzymatic function is known) and *putative enzymes* (i.e., enzymes with weak similarity to sequences of known function). From these initial "hits", within a phylogenetic perspective, we identify an initial set of pathways. This set can be used to generate a set of *expected enzymes* (i.e., enzymes that have not been clearly detected, but that would be expected given the set of hypothesized pathways) and *missing enzymes* (i.e., enzymes that occur in the pathways but for which no sequence has yet been biochemically identified for any organism). Further reasoning identifies tentative connective pathways and *necesssary pathways*, as follow from growth medium requirements.

*Work supported in part by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy under contract XX-XX-NN-NNNNN.

1 T. Gaasterland and C. Sensen, MAGPIE: A Multipurpose Automated Genome Project Investigation Environment for Ongoing Sequencing Projects. In *Bacterial Genomes: Physical Structure and Analysis*, ed. G. Wienstock et al. (to appear).

2 T. Gaasterland, J. Lobo, N. Maltsev, and G. Chen. Assigning Function to CDS Through Qualified Query Answering. In *Proc. 2nd Int. Conf. Intell. Syst. for Mol. Bio.*, Stanford U. (1994).

3 T. Gaasterland and E. Selkov. Automatic Reconstruction of Metabolic Structure from Incomplete Genome Sequence Data. In *Proc. Int. Conf. Intell. Syst. for Mol. Bio.*, Cambridge, England (1995).

3DBase - A Macromolecular Structure Database

E. E. Abola, J. Prilusky, N. O. Manning, J. L. Sussman, Protein Data Bank, Department of Chemistry, Brookhaven National Laboratory, Upton, NY 11973 and Bioinformatics Unit, Weizmann Institute of Science, 76100 Rehovot, Israel.

The Protein Data Bank (PDB) is an archive of experimentally-determined three-dimensional structures of proteins, nucleic acids, and other biological macromolecules. PDB has a 25-year history of service to a global community of researchers, educators, and students in a variety of scientific disciplines. The common interest shared by this community is the desire to access information that can relate the biological functions of macromolecules to their three-dimensional structures. We now report the construction of a new relational database, 3DBase, that provides access to knowledge and information on macromolecules using a high-level query language.

The complexity of PDB entries and their use by a multi-disciplinary community required the construction of a database that represents structural, biological, chemical, and bibliographic information. In addition to all coordinate entries found in PDB, the database contains semantic links to entries found in other databases. For example, 3DBase represents the relationships between sequences found in PDB with those in SWISSPROT, GSDB, or GenBank.

3DBase uses Victor Markowitz's Object Protocol Model (OPM) and the SYBASE DBMS engine. OPM's object-oriented view provides a scientifically intuitive representation of the data while SYBASE provides a powerful and robust environment for data management. Two primary objects in 3DBase are oExperiment and oMacroMolecule. These objects describe the experiment and the biologically active molecule, extending the current view found in PDB entries.

Database interoperability is addressed through the use of schema sharing and support for a variety of data interchange format in query results. 3DBase uses the CitDB schema developed by GDB to store literature references. In the near future the CitDB at PDB will be merged with GDB's data, thus making available a single CitDB database containing all the references of interest to the genomic community. In addition, 3DBase uses similar base class objects found in GDB. An example is GDB's powerful and elegant solution to the problem of providing user-supplied annotation to individual objects in the database.

Access to 3DBase is primarily through a Web browser constructed using the Genera software package developed by Stan Letovsky. In addition to accessing data stored in 3DBase, the browser provides links to entries in other databases. Graphical views of molecules are provided in the browser by use of R. Sayle's Rasmol program along with graphical annotation commands stored in the database. Access to 3DBase *via* SQL or OPM's QLT language will also be made available to those wishing to pose complex queries not available through the browser.

* The Protein Data Bank is supported by funds from the U. S. National Science Foundation, the U. S. Public Health Service, National Institutes of Health, National Center for Research Resources, National Institute of General Medical Sciences, National Library of Medicine, and the U. S. Department of Energy under contract DE-AC02-76CH00016.

TIGR DATABASE (TDB): INTEGRATED BIOLOGICAL DATABASES TO SUPPORT RESEARCH IN GENE EXPRESSION, PROTEIN FAMILIES AND GENOME EVOLUTION

Anthony R. Kerlavage, Mark D. Adams, Judith Blake, Carol Bult, Rebecca Clayton, Lisa FitzGerald, Anna Glodek, Michael Heaney, Robert Shirley, Granger Sutton, Owen White, and J. Craig Venter, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850

The recent determination of the complete sequence of the genomes of *Haemophilus influenzae*[1] and *Mycoplasma genitalium*[*2] have provided a tremendous amount of new information about genome organization and evolution on a whole-genome scale. In addition, the application of random sequencing from a large number of cDNA libraries to produce expressed sequence tags (ESTs) has provided a wealth of new information about gene expression, protein families and evolution[3]. The existence of these global datasets allows new kinds of questions to be posed; however, current data banks are inadequate to support the types of queries scientists will want to make using these data. We have designed the TIGR Database (TDB), a collection of databases of sequences and related data, to handle these requirements.

TDB currently consists of several components available via the World Wide Web (URL: http://www.tigr.org/tdb/tdb.html). The TIGR Microbial Database (MDB) provides curated information on complete bacterial genomes sequenced at TIGR, including putative genes with cellular role classifications. The Human cDNA Database (HCD) provides access to the most complete set of available EST data and offers name, role and sequence searching capability. The Expressed Gene Anatomy Database (EGAD) provides curated biological information relating to the function, role, expression and isology of cDNAs and their corresponding proteins. Sequences, Sources, Taxa (SST) integrates DNA and protein sequence data with specimen, collection and taxonomic information. All databases in TDB have been designed to facilitate interoperability with other relational databases of biological information, including the other TDB databases as well as external DNA, protein, mapping and citation databases.

*Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy under contract DE-FC02-95ER61962.A001.

[1]Fleischmann *et al., Science* **269**, 449-604 (1995).
[2]Fraser *et al., Science,* in press (1995).
[3]Adams *et al., Nature,* **377 suppl.**, 3-174 (1995).

DATA MANAGEMENT FOR THE RESOURCE FOR MOLECULAR CYTOGENETICS *

Manfred D. Zorn and Jenny E. Marstaller, Software Technologies and Applications Group, Information and Computing Sciences Division, Lawrence Berkeley National Laboratory, University of California, Berkeley CA 94720.

The LBL/UCSF Resource for Molecular Cytogenetics has been created to facilitate the application of molecular cytogenetics in clinical and biological studies. Work is being pursued in three areas: Development and application of improved hybridization technology, selection of probes optimized for use in fluorescence in situ hybridization (FISH) and development of digital imaging microscopy. All of these areas entail creation and manipulation of large images and other laboratory data. Our group is focussed to provide data management support for all the activities in the Resource.

To facilitate the free data exchange between researchers at UCSF and LBL which are a few miles apart we developed a Mosaic interface to access and modify information using the World Wide Web. The data are located on a central database. The Mosaic client allows to formulate retrieval and edit operations that are sent to the database. Results are filtered through a Perl script which generates HTML documents with Hypertext links that are sent back to the Mosaic client. Data from the Resource are made available using a similar mechanism that is open to outside access.

Probe information and mapping data from the Resource are being submitted to public databases, i.e., GDB. In a collaboration with GDB we have developed a data submission tool (see separate abstract by Manfred Zorn) to facilitate the distribution of our research results.

In order to handle large amounts of images we are developing an image annotation database. The images themselves are automatically transferred to the LBL Mass Storage System. The annotation will be reformatted and loaded into a relational database to allow efficient query processing.

We will present an overview and the current status of our work.

* This work was supported by the Director, Office of Energy Research, Office of Health and Environmental Research, Human Genome Program, of the US Department of Energy under Contract No. DE-AC03-76SF00098.

A DATABASE SYSTEM TO SUPPORT GENOME MAPPING

Mark Graves, Dept. of Cell Biology, Baylor College of Medicine, Houston, TX 77030.

We have developed a small database system to support physical mapping in genome research laboratories. A database system consists of a collection of data (called a database) and software which manipulates the data (called a database management system). To support genome mapping we have included other software which simplifies the development of genome databases. The software includes: a schema design tool, a data entry tool, and a query-by-example tool. Additional mapping-specific tools have been developed, including a tool to record the result of filter hybridization experiments.

Our database management system is based on the storage of graphs as binary relationships. A graph is a collection of vertices, edges, and labels which form binary relationships. A binary relationship is used to describe one attribute of a entity, such as the name of a person. An advantage of binary relationships is that they simplify the representation of data to a form which is easier to manipulate. Binary relationships are stored and retrieved using our graph database management system [1,2,3].

A database schema is a description of the data in a database. The database developer uses our schema design tool to draw the types of data to be included in the database. The schema is a collection of graph templates, and each template defines the binary relationships which are to be created for each data type. For example, a "person" graph template might create binary relationships for the name, phone number, and email address of a person.

Our data entry tool uses the graph templates to create a graphical user interface in which the user enters data. The data entry forms are created automatically from the templates in the schema, and if the schema changes the data entry tools change, too. The automatic creation of data entry forms allows the database to be rapidly developed and to be modified when the laboratory process changes: something which is common in the rapidly changing field of genetics.

Query-by-example is a database query paradigm where the user enters a query by specifying part of the data in a template, and the query tool fills in the template based on the data in the database. Our query-by-example tool creates a graphical user interface for each graph template, as is done in the data entry tool. The user enters data into part of the template, and the system generates a report of all the data which matches the partially filled in template.

These three tools form the core of our system, but additional tools can be added to simplify part of the process. One tool is to record filter hybridizations using a graphic display. Instead of using the Data Entry Tool to enter data textually about hybridizations, the user can enter hybridization data using a graphical display corresponding to the filter. This is currently being used at the Baylor College of Medicine Human Genome Center to record cosmid filter hybridization experiments for human chromosome X and 17.

* Supported by a DOE Human Genome Distinguished Postdoctoral Fellowship.

1. M Graves, ER Bergeman, CB Lawrence. "A Graph-Theoretic Data Model for Genome Mapping Databases". In Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences, Vol 5, pp 32-41. IEEE Press. January, 1995.

2. M Graves, ER Bergeman, CB Lawrence. "Graph Database Systems". IEEE Engineering in Medicine and Biology special issue on Genomics. November, 1995.

3. M Graves, ER Bergeman, CB Lawrence. "Querying a Genome Database Using Graphs". In Proceedings of The Third International Conference on Bioinformatics and Genome Research. HA Lim and CR Cantor, eds. World Scientific Publishing Co., Singapore. 1995. (in press)

A FLEXIBLE DATABASE SCHEMA FOR LARGE-SCALE PHYSICAL MAPPING AND SEQUENCING ACROSS MULTIPLE GENOMES

Tom Slezak (slezak@llnl.gov), Mark Wagner, T. Mimi Yeh. Human Genome Center, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94550

The Human Genome Center at LLNL has developed a relational database over the last 6 years to support our work on chromosome 19. We intentionally designed the database to support our specific physical mapping needs. We anticipated applying what we learned form this experience when we expanded our capabilities to support new approaches and to cover other genetic real estate.

Effective closure of the physical map of ch19 has been declared and we are now retargeting our efforts towards production sequencing (and associated high-resolution mapping) of certain gene families across the entire human genome, regions of interest in other genomes with conserved synteny with respect to humans, and potential work on various bacterial, plant, and animal genomes. Separate databases for each chromosome or genome are not well-suited for the comparative biology that lies in our future. We must scale up our database to be able to handle queries on mapping and sequencing data that span all our targets regardless of species, provide for better public access to data, and fully participate in the Federation of Genome Databases.

At the last meeting we reported on our design ideas. Major concepts include: all objects have a unique, permanent identifier; objects and relations will be highly abstracted (single base-class tables for all clones, probes, hybridization results, etc.); each object and relation can be flagged as public or private; object storage handled separately, etc. Key to this design is a central global identifier table, which tracks information on every object and relation, providing a flexible and simple reference interface mechanism. This design stems from the generic "mappable object" abstraction which allowed us to successfully integrate our physical mapping data, as described at earlier meetings.

This database has been implemented and our 250Mb of ch19 data transferred to it, using a database re-engineering tool (described by Mark Wagner, et. al. in a separate poster.) We are currently implementing WWW interfaces to this database for data entry and non-graphical querying and have modified our graphical browser to read from it. We will discuss performance implications of this abstracted schema design from our early experience with it and extrapolate on its ability to scale to meet our future needs.

(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.)

THE BCM/UH NETWORK SERVER CORE AND SEARCH LAUNCHER

Daniel B. Davison1,3, Brent A. Wiese2, Istvan Ladunga2, Kim C. Worley2, and Randall F. Smith1,2 1 Department of Cell Biology, 2 Human Genome Center and Department of Molecular and Human Genetics, Baylor College of Medicine, and 3 University of Houston, Houston TX.

We are providing a variety of molecular biology-related search and analysis services to Genome Program investigators to improve the identification of new genes and their functions. These services are available via the BCM Search Launcher World Wide Web (WWW) pages which are organized by function and provide a single point-of-entry for related searches. Pages are included for 1) protein sequence searches, 2) nucleic acid sequence searches, 3) multiple sequence alignments, 4) pairwise sequence alignments, 5) gene feature searches, 6) sequence utilities, and 7) protein secondary structure prediction. The Protein Sequence Search Page, for example, provides a single form for submitting sequences to WWW servers that provide remote access to a variety of different protein sequence search tools, including BLAST, FASTA, Smith-Waterman, BEAUTY, BLASTPAT, FASTAPAT, PROSITE, and BLOCKS searches. The BCM Search Launcher extends the functionality of other WWW services by adding additional hypertext links to results returned by remote servers. For example, links to the NCBI's Entrez database and to the Sequence Retrieval System (SRS) are added to search results returned by the NCBI's WWW BLAST server. These links provide easy access to Medline abstracts, links to related sequences, and additional information which can be extremely helpful when analyzing database search results. For novice, or infrequent users of sequence database search tools, we have pre-set the parameter values to provide the most informative first-pass sequence analysis possible.

A batch client interface to the BCM Search Launcher for Unix and Macintosh computers has also been developed to allow multiple input sequences to be automatically searched as a background task, with the results are returned as individual HTML documents directly on the user's system. The BCM Search Launcher as well as the batch client are available on the WWW at URL **http://gc.bcm.tmc.edu:8088/ search-launcher/launcher.html**.

The BCM/UH Server Core provides the necessary computational resources and continuing support infrastructure for the BCM Search Launcher. The BCM/UH Server Core is composed of three network servers and currently supports electronic mail, and WWW-based access; ultimately, specialized client-server access will also be provided. The hardware used includes a 2048-processor MasPar massively parallel MIMD computer, a DEC Alpha AXP/OSF1, a Sun 2-processor SparcCenter 1000 server, and several Sun Sparc workstations.

In addition to grouping services available elsewhere on the WWW, and providing access to services developed at BCM and UH, the BCM/UH Server Core will also provide access to services from developers who are unwilling or unable to provide their own Internet network servers. One such service from Dr. Don Gilbert, "GenBank Subset Search", will allow one to search a user-specified subset of GenBank via email and the WWW. Another tool under development is The Institute for Genomic Research's multiple sequence alignment program (MSA) which uses simulated annealing.

A major focus for future development is our collaboration with the Genome Sequence Data Base (GSDB) to allow analysis services provided by the Server Core to be integrated within the GSDB Annotator by inter-process communication. This will allow a researcher to perform in-depth sequence analysis and easily integrate the results into sequence database annotations.

This research is supported by grants to D.D. from the U.S. Department of Energy Office of Health and Environmental Research (DE-FG03-95ER62097/A000), the National Library of Medicine (1R01-LM05792), the National Science Foundation (BIR 91-11695), a National Research Service Award to K. C. W. (1F32-HG00133-01), a grant to the Baylor Human Genome Center (P30-HG00210), and a grant to R. F. S. (1R01-HG00973-01) from the National Center for Human Genome Research, National Institutes of Health.

ENHANCED RESOURCES FOR ACCESSING GENETIC AND BIOLOGICAL INFORMATION VIA THE WORLD WIDE WEB*

Amy K. Voltz and Kenneth H. Fasman, Division of Biomedical Information Sciences, Johns Hopkins University School of Medicine, Baltimore, MD 21205-2236

The World Wide Web (WWW, the "web") provides a user-friendly interface for finding and retrieving genomic data in the Genome Data Base (GDB). Current web access via the "GDB Browser" allows the user to access all data objects in GDB, but does not emphasize data that may be most relevant to biologists interested in specific information regarding genes.

We have developed a new interface which directs the user to particular information about genes. This interface has been designed with specific questions in mind, the answers to which can now be easily found in the database: Is this gene in GDB? Where is the gene located in the genome, and what other genes or markers are located nearby? Are there PCR primers or genomic/cDNA clones for this gene? Who else has information about this gene?

These questions can be asked and answered with ease using the new interface, which returns the requested information in concise, formatted tables. The user also has the option of viewing detailed information for any object, by following a highlighted link from this table via GDB accession number.

In addition to the human gene mapping information in GDB, there are many databases currently in existence that store information of relevance to biologists. These include databases of nucleotide sequences (GSDB, GenBank, EMBL, DDBJ), human genetic disease (OMIM), protein sequence and structure (Swiss-Prot, PIR, PDB), and the genomes of model organisms (MGD, FlyBase).

Although the Internet and World Wide Web provide access to many of these databases and include some links that enable users to move from one database to another, most of these data must be gathered by contacting multiple databases. The databases do not provide an integrated view of biological knowledge -- rather the information is presented as individual entries in each database.

We have developed a prototype system for the integration of biological information. Entries in the Gene Family Database (http://gdbdoc.gdb.org/~avoltz/home.html) begin with a definition of a gene family and descriptions of its members, including links to the databases that compile proteins of similar sequence and motif into functional families (PRODOM, PROSITE, BLOCKS). The entries also include data on map location, nucleotide sequence, gene structure and function, RNA transcripts, protein sequence, structure and expression. There are also hypertext links to information regarding model organisms and human genetic disease.

Future entries are being developed with collaborators, who can serve to analyze the database information for accuracy, but more importantly will help to develop the content and presentation of the information in the Gene Family Database. We are working to develop a model for community-based curation of such a resource.

*Supported in part by the National Institutes of Health, National Research Service Award number 1F32 HG00148-01 from the National Center for Human Genome Research, and Department of Energy award number DE-FC02-91ER61230.

A Graphical Ad Hoc Query Interface Capable of Accessing Heterogeneous Public Genome Databases

Joseph Leone[1]

A. J. Boggs & Company
2853 W. Jolly Road
Okemos, Michigan 48864

Dong-Guk Shin[2]

Computer Science & Engineering.
University of Connecticut
Storrs, CT 06269-3155

Interoperability with public genome databases is expected to be crucial in making the Human Genome Project a success. We propose to develop software tools in which users of the genome community can learn and/or examine public genome database schemas in a relatively short time and can produce a correct SQL expression easily. Specifically we aim at developing tools in which users access both GDB and GSDB simultaneously.

Initially we will focus on addressing one of the known key problems for making a distributed SQL interface practically useful. That is, a user who is not familiar with one of the two public genome databases would have difficulty understanding the schema of that database. Consequently, he may not be able to import relevant database schemas among many available ones and may have difficulty forming correct SQL expressions. In a federated database environment, this problem of dealing with unfamiliar third party database schemas becomes much more severe due to the manifold added complexity. Our proposed approach is to design features that aid users in understanding GDB and GSDB schemas quickly and embed them as a part of the proposed distributed SQL interface.

Another aspect of our effort focuses on testing the feasibility of using Galaxy graphical user interface development tool kit to build the proposed interface. Our ultimate goal is to make the interface system portable across multiple hardware platforms, i.e., Unix workstations, PCs and Macs. The proposed interface will be a client program, and by developing the client program to be portable across different platforms, the interface will be usable by the widest possible user groups.

1. Work supported by a grant from the U.S. Department of Energy, under Contract No. DE-FG02-95ER81906.

2. Work done in collaboration with Genome Data Base, Johns Hopkins University, and supported by National Center for Human Genome Research, National Institute of Health, HG00772-01.

On-Demand Integration of Biological Data [1]

J. Crabtree, L. Wong, P. Buneman, S.B. Davidson, and C. Overton
Dept. of Computer and Information Science & Dept. of Genetics
University of Pennsylvania, Philadelphia, PA 19104
Email: {crabtree,limsoon,peter,susan,coverton}@cis.upenn.edu

We have implemented a general-purpose query system, CPL/Kleisli, that provides access to a variety of "unconventional" data sources (e.g., ACeDB, ASN.1, BLAST), as well as to "standard" relational databases. The system represents a major advance in the ability to integrate the growing number and diversity of biology data sources conveniently and efficiently. It features a uniform query interface across heterogeneous data sources, a modular and extensible architecture, and most significantly for dealing with the Internet environment, a programmable optimizer. We have demonstrated the utility of our system in composing and executing queries that were considered difficult, if not unanswerable, without first either building a monolithic database or writing highly application-specific integration code (details and examples available at `http://agave.humgen.upenn.edu/cpl/cplhome.html`). In conjunction with other software developed in our group, we have assembled a toolset that supports a range of data integration strategies as well as the ability to create specialized databases initialized from community databases (see abstracts by Buneman et. al and Davidson et. al in this meeting). Our integration strategy is based upon the concept of "mediators", which serve a group of related applications by providing a uniform structural interface to the relevant data sources. This approach is cost-effective in terms of query development time and maintenance. This approach is cost-effective in terms of query development time and maintenance. Here we discuss recent results in optimizing queries such as "retrieve all known human sequence containing an Alu repeat in an intragenic region" where the data sources are heterogeneous and distributed across the Internet.

CPL already optimizes queries in such a way as to minimize response time, and it is open to the addition of both new data sources and algebraic rules governing the use of those data sources. To determine how best to augment CPL's query transformation rules, we have conducted a series of performance tests across different combinations of biological data sources. For the test query "retrieve all sequence entries with a CDS feature located on chromosome A," a CPL query spanning GDB and GSDB approaches the response time of the query executed directly on GSDB, which maintains a local copy of the necessary GDB information. In contrast, the best version of the same query executed across GDB and NCBI-Entrez using our ASN.1 query engine is at least an order of magnitude slower. The tests allowed us to identify optimization rules which apply to large classes of queries and are hence reusable.

To aid in the iterative process of identifying potential bottlenecks and introducing rules to circumvent them, we are developing a set of graphical profiling tools. One such tool displays the alternative query plans generated by the system and a second monitors the actual execution of a query plan, displaying the pattern of data source accesses generated by the system. Profile analysis has enabled us to identify which of the data retrievals were dominating the time spent on our test queries, much as a programming language profiler can reveal how much time is being spent in a specific subroutine or loop. The difference in our case is that the time taken to fetch a particular piece of data is in general dependent on many more variables (network traffic, remote server usage, and so on).

We have found that for CPL (or an analagous system) to decide which optimization strategy to employ in a given situation requires access to meta-data pertaining to the data sources it accesses. Extending the current system to be aware of such information where it is available will bring a twofold advantage. On the one hand, such information is almost certain to be essential in arbitrating between different optimization rules and classes of rules. On the other hand, since there is often overlap between what the optimizer needs to know to generate an efficient plan and what a user needs to know to compose a query, an obvious extension is to enable the system to guide a user's query based on its (necessarily) up-to-date knowledge of the data sources. A competent query interface should serve not only to hide irrelevant details, but also to provide relevant details. Thus our two immediate goals–usability and efficiency–are not necessarily orthogonal, as they might first appear, and we hope to exploit the connection.

[1] This research was supported by a grant from the Director, Health Effects and Life Science Research Division, Office of Health and Environmental Research of the U.S. Department of Energy under contract DOE DE-FG02-94-ER-61923 Sub 1.

GRAIL-genQuest: A Comprehensive Computational Framework for DNA Sequence Analysis[1]

Ruth Ann Manning and Björn E.F. Mossberg, ApoCom Inc., 1020 Commerce Park Drive, Suite F, Oak Ridge, TN 37830-8026.

Since 1992[2] the **GRAIL** DNA sequence analysis and companion **genQuest**[3] database comparison systems developed at Oak Ridge National Laboratory (ORNL) have been publicly available over the Internet. Because of the potential for security and patent compromise over the Internet, these versions are not currently available to many researchers in pharmaceutical and biotechnology companies who cannot send proprietary sequences past their data-secure firewalls. ApoCom is providing the **GRAIL-genQuest** software to domestic and international customers to run self-contained over their own local area networks.

Although DNA sequencing in the Human Genome Project is occurring fairly systematically, biotechnology companies have focused on sequencing regions thought to contain particular disease genes. **GRAIL**, the most accurate and widely used computer-based system for locating and characterizing genes in DNA sequences, is not accessible to many of them because their primary computers are Macintoshes and personal computers (PCs). Since ORNL has developed client tools only for high-end UNIX-based computer workstations, ApoCom is developing a prototype cross-platform client graphical user interface (GUI) for **GRAIL-genQuest** during Phase I.

The growth of genome databases is expected to continue at a fast pace in the attempt to sequence the human genome completely by the year 2005. Parallel processing will be a viable solution to handle searching through this ever-increasing volume of data. During Phase I **genQuest** database searching algorithms will be parallelized for both shared- and distributed-memory computer platforms.

Prototype graphical interface systems for Macintosh™, NT Windows™, and Windows 95™ that mimic the function and operation of the current **GRAIL-genQuest** clients will enable a larger portion of biotechnology companies to make use of the **GRAIL** suite of analysis tools. Parallel **genQuest** servers will improve response time for searches and increase user capacity per server. Such fast shared- and distributed-memory computing solutions using general-purpose hardware will improve the price-performance ratio and make parallel searches more affordable to the biotechnology community.

---

[1] Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy under contract DE-FG02-95ER81923.

[2] E.C. Uberbacher and R.J. Mural, *Locating protein coding regions in human DNA sequences by a multiple sensor neural network approach,* Proc. of National Academy of Sciences, USA, Vol. 88, 1991, pp. 11261-11265.

[3] E.C. Uberbacher, R.J. Mural, X. Guan, S.Petrov, and M.B. Shah, *genQuest: A Sensitive Sequence Comparison Server for DNA and Proteins*, Genome Mapping and Sequencing, Cold Spring Harbor, NY, May 11-15, 1994.

A TOOL FOR REENGINEERING LARGE DATABASES

Mark C. Wagner (mwagner@llnl.gov), Shannon Waller, T. Mimi Yeh, and Thomas R. Slezak.
Human Genome Center, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94550

The Human Genome Center at Lawrence Livermore National Laboratory has been developing and using a relational database over the last 6 years. This database was designed to support our immediate need to hold only our data for human chromosome 19. We have accomplished our physical mapping task for chromosome 19 and are now poised to map and sequence other areas of human and non-human genomes. Since our current database is human chromosome 19 specific, a new database schema was required to accomplish these goals (see the poster by Tom Slezak, et. al. for details). This schema uses a large degree of abstraction and different ways of organizing the data in the original database, reducing the table count from 200 to approximately 150.

Reengineering a large database and populating it with the contents of the original is a complex task, certain to require many iterations. We needed a tool specific to this task that would let us specify all details of the new database in special "meta tables", plus routines which could automatically generate all the table, rule, trigger, index, user, and permission SQL. A byproduct of this tool would be on-line and current documentation, as well as table-specific backup dump procedures. We have also had to devise a method to "translate" our data from our existing database into our new schema. This is done at column-level granularity: for each column in the new database, the SQL necessary to fill that column is specified. In this fashion, if a column is moved or reordered, a simple automated routine can readily extract the data for uploading into the new table. Changes to the schema are effected in the meta tables. The process of generating and populating the new database from the old is completely automated.

We used PERL and HTML to create a WWW graphical interface to develop this translation tool. This permitted us to develop one interface for use across multiple platforms with a minimum of effort.

It should be noted that this tool was not designed for implementing a database from scratch, although it could be used in that application. The strength of this tool lies in its ability to move data from one database schema to another with a minimum amount of human intervention.

(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.)

SUBMITDATA: DATA SUBMISSION TO PUBLIC GENOMIC DATABASES*

Manfred D. Zorn, Software Technologies and Applications Group, Information and Computing Sciences Division, Lawrence Berkeley National Laboratory, University of California, Berkeley CA 94720

Making information generated by the various genome projects available to the community is very important for the researcher submitting data and for the overall project to justify the expenses and resources. Public genome databases generally provide a protocol that defines the required data formats and details how they accept data, e.g., sequences, mapping information. These protocols have to strike a balance between ease of use for the user and operational considerations of the database provider, but are in most cases rather complex and subject to change to accommodate modifications in the database.

SubmitData is a user interface that formats data for submission to GSDB or GDB. The user interface serves data entry purposes, checking each field for data types, allowed ranges and controlled values, and gives the user feedback on any problems. Besides one-time submissions, templates can be created that can later be merged with TAB-delimited data files, e.g., as produced by common spreadsheet programs. Variables in the template are then replaced by values in defined columns of the input data file. Thus submitting large amounts of related data becomes as easy as selecting a format and supplying an input filename. This allows easy integration of data submission into the data generation process.

The interface is generated directly from the protocol specifications. A specific parser/compiler interprets the protocol definitions and creates internal objects that form the basis of the user interface. Thus a working user interface, i.e., static layout of buttons and fields, data validation, is automatically generated from the protocol definitions. Protocol modifications are propagated by simply regenerating the interface.

The program has been developed using ParcPlace VisualWorks and currently supports GSDB, GDB and RHdb data submissions. The program has been updated to use VisualWorks 2.0. We will present an overview and the current status of our work.

* This work was supported by the Director, Office of Energy Research, Office of Health and Environmental Research, Human Genome Program, of the US Department of Energy under Contract No. DE-AC03-76SF00098.

PROVIDING OBJECT-ORIENTED INTERFACES FOR MOLECULAR BIOLOGY DATABASES*

Victor M. Markowitz, I-Min A. Chen, Anthony Kosky, and Ernest Szeto, Information and Computing Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

Numerous repositories for molecular biology data are implemented using commercial relational database management systems (DBMSs). These DBMSs provide reliable facilities for managing data in molecular biology databases (MBDs) such as the Genome Sequence Data Base (GSDB), but do not provide constructs for directly representing application-specific objects, such as sequences and alleles: such objects are usually represented by disconnected tuples that are scattered among multiple tables. Interacting with MBDs implemented using relational DBMSs can be substantially simplified by providing these MBDs with object-oriented interfaces for examining (browsing and querying) their structure and content, and thus insulating users and applications from the underlying DBMSs.

The first step of developing an object-oriented interface for a relational MBD is constructing an object view for the MBD. For most MBDs, constructing such a view cannot be carried out automatically. It is often hard to determine algorithmically what tables represent classes of objects and what tables represent attributes (e.g., set-valued attributes) that cannot be represented as columns in the tables representing classes. Furthermore, the foreign-key information which is essential for determining the relationships between different classes of objects is often missing in relational MBD definitions. Consequently, an interactive procedure is needed for filling in the missing information and/or determining the structure of the desired object view for a relational MBD.

We have developed a retrofitting tool that allows constructing and maintaining object views on top of existing relational MBDs, without affecting the structure and content of MBDs, and therefore without disturbing existing applications based these MBDs. This tool is based on the Object-Protocol Model (OPM). OPM provides constructs for modeling objects and protocols (laboratory experiments) specific to molecular biology applications [1]. The OPM retrofitting tool generates a canonical OPM view using all the available information on the relational MBD, and provides facilities for interactively refining the OPM view by renaming classes and attributes, changing the value classes of attributes, hiding classes and attributes from the view, defining new derived classes and attributes, grouping simple attributes into tuple attributes, adding and removing classes and attributes, merging and splitting classes, and so on.

Once an OPM view has been developed for a relational MBD, the object-oriented interface on top of the MBD is provided by the OPM data management tools. The OPM retrofitting tool has been applied for constructing an OPM view for Genome Sequence Data Base (GSDB) 2.2 in order to allow using the OPM browsing and querying tools on top of GSDB 2.2.

*Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research, of the U.S. Department of Energy under Contract DE-AC03-76SF00098.

[1] Chen, I.A., and Markowitz, V.M., An Overview of the Object-Protocol Model (OPM) and OPM Data Management Tools, *Information Systems*, Vol 20, No 5 (July 1995), pp. 393-418.

VERSION 4 OF THE OPM DATA MANAGEMENT TOOLS: ENHANCED SUPPORT FOR MOLECULAR BIOLOGY DATABASES*

Victor M. Markowitz, I-Min A. Chen, Ernest Szeto, and Jia-Lin N. Chen, Information and Computing Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

Commercial relational database management systems (DBMSs) provide data management facilities that are essential for operating large production molecular biology databases (MDBs), such as Genome Data Base (GDB), Genome Sequence Data Base (GSDB), and Protein Data Bank (PDB). However, developing and maintaining large MBDs with relational DBMSs are complex, error-prone, and time-consuming processes. Furthermore, the large, low-level, DBMS-specific relational definitions of such MBDs is almost incomprehensible to scientists.

The Object-Protocol Model (OPM) provides scientists with high-level, concise, DBMS-independent languages for specifying the structure of and manipulating data in MBDs. We have developed a suite of data management tools based on OPM, including a graphical OPM schema editor, an OPM to DBMS schema translator, an OPM based data entry and query tool, an OPM retrofitting tool, and an OPM data loading utility. For MBDs developed with relational DBMSs, the OPM data management tools substantially improve the efficiency of developing, maintaining, and interacting with (e.g., querying and browsing) MBDs [1].

The OPM data management tools are currently used for developing several new MBDs, such as the new versions of GDB and PDB, and for providing object-oriented interfaces on top of existing MBDs, such as GSDB. Interactions with the GDB, PDB, and GSDB groups over the past year revealed the need for extending the OPM tools with new features. Accordingly, version 4 of the OPM data management tools provide facilities for: (1) controlled vocabularies consisting of terms that are codified and associated with detailed descriptions; (2) object versions representing alternative experimental and analysis data and recording historic information; (3) cross-referencing MBDs in order to facilitate molecular biology data exploration across multiple databases; (4) different strategies for querying efficiently and updating MBDs; (5) interactively retrofitting OPM schemas on top of existing MBDs; (6) constructing customized interfaces for browsing, querying, and updating MBDs, via an API; and (7) physical database design (e.g., indexing, segment allocation) for improving the efficiency of data manipulation in large MBDs.

Current work on the OPM data management tools includes developing new OPM tools that will provide facilities for (1) restructuring MBDs as a result of structural changes entailed by the evolution of their underlying applications, and (2) developing MBDs with object-oriented DBMSs.

The OPM tools, documentation, and papers are available via World Wide Web using URL: `http://gizmo.lbl.gov/opm.html`.

*Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research, of the U.S. Department of Energy under Contract DE-AC03-76SF00098.

[1] Chen, I.A., and Markowitz, V.M., An Overview of the Object-Protocol Model (OPM) and OPM Data Management Tools, *Information Systems*, Vol 20, No 5 (July 1995), pp. 393-418.

QUERYING HETEROGENOUS MOLECULAR BIOLOGY DATABASES*

Victor M. Markowitz, I-Min A. Chen, Anthony Kosky, and Ernest Szeto, Information and Computing Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

Molecular biology data are scattered among multiple data repositories, including molecular biology databases (MDBs). Although containing related data, these repositories are often isolated and are characterized by various degrees of heterogeneity: they usually represent different views (schemas) of the molecular biology domain and are implemented using different database management systems (DBMSs). Comprehensive studies of biological data often involves examining data across heterogeneous databases.

Solutions currently promoted for querying data across heterogenous MBDs involve constructing MBD *federations* or *data warehouses*, such as the Genome Topographer (Cold Spring Harbor Laboratory) and the Integrated Genomic Database (German Cancer Research Institute). These solutions entail construction of a *global* view of a collection of MBDs, where definitions of the component MBDs are expressed in a common language and discrepancies between these definitions are resolved before they are integrated into a global view. For data warehouses, data from MBDs must be also loaded into a central data repository. The main problem of MBD federations and data warehouses is the complexity of constructing global views. Data warehouses have also the additional problems of not being synchronized with evolving component MBDs and of potentially extremely large physical sizes.

Querying heterogenous MBDs can be achieved without constructing MBD federations or data warehouses, by organizing MBDs in a loose *multidatabase* system. We have developed a multidatabase query strategy for MBDs implemented using relational DBMSs, in the context of the Object-Protocol Model (OPM) data management tools [1]. For MBDs that have not been developed using OPM, OPM views of the MBDs are first constructed using an OPM retrofitting tool. Then, existing OPM tools provide facilities for examining MBD schemas and browsing and querying individual MBDs associated with OPM views.

Our multidatabase query strategy is based on an MBD dictionary that contains information on MBDs, including their OPM views, DBMS implementation, and links to other MBDs. A multidatabase query tool processes queries over heterogenous MBDs associated with OPM views, by (1) decomposing these queries into subqueries for individual MBDs, (2) using exiting OPM query tools for processing the subqueries, and (3) assembling subquery results into multidatabase query results. Our query strategy assumes that users understand the structure and semantics of the MBDs they query. In a related project, we plan to develop an MBD Library containing comprehensive documentation on MBDs and with facilities that will assist users in expressing multidatabase queries.

Work is underway on applying the multidatabase query strategy outlined above for supporting queries over the new versions of Genome Data Base (GDB), Genome Sequence Data Base (GSDB), and Protein Data Bank (PDB).

*Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research, of the U.S. Department of Energy under Contract DE-AC03-76SF00098.

[1] Chen, I.A., and Markowitz, V.M., An Overview of the Object-Protocol Model (OPM) and OPM Data Management Tools, *Information Systems*, Vol 20, No 5 (July 1995), pp. 393-418.

SPACE: A Flexible Database and Analysis Tool For Genomic Sequence Reconstruction and Multi-Species Mapping

Arun Aggarwal, Sam Pitluck, Frank Eeckman
Human Genome Center Informatics Group,
Lawrence Berkeley National Laboratory,
University of California, Berkeley, CA 94720

SPACE, or Sequencing Platform using ACE, is a variant of ACeDB, a suite of database, analysis, and display software originally developed by Richard Durbin and Jean Thierry-Mieg to meet the needs of the C. elegans genome research community. SPACE extends the functionality of ACeDB and builds on SynDB, an earlier LBNL variant of ACeDB used primarily for human and mouse mapping data.

Now being used in our production laboratories, SPACE's purpose is to provide additional tools to meet the evolving requirements of LBNL's sequencing and mapping projects. It differs from its predecessors primarily in the addition of a new display module which allows the simultaneous viewing of multiple maps within the same window and in the addition of a new sequence assembly and editing tool.

For our directed sequencing effort, the assembly/editing tool is used to view pre-assembly mapping data, visually select sequence fragments, transparently set any assembly constraints, call an assembler on the selected fragments, view the results, and edit wherever necessary. Furthermore, with the flexibility of the multi-map display, different assemblies can be compared side-by-side or a given assembly can be viewed with respect to higher level mapping data. For these assemblies, we are using SPASS, a C-level routine which incorporates the Fragment Assembly Kernal (FAK) written by Gene Myers, Susan Larson, and Mudita Jain (please refer to an accompanying LBNL poster on SPASS).

For our mapping efforts, the physical map display allows the viewing of loci and clones on the chromosome. And, using the multi-map, different maps, either of the same region or of different regions from different species, can be displayed. To aide in cross-species analysis, new data structures have been added to display direct homology comparisons between the species.

SPACE uses ACeDB's basic data structures, which allow all parts of the database to be easily cross-referenced, and ACeDB's user interface, which permits exploration of data via "point and click" with the computer mouse. Other ACeDB modules allow users to display and search DNA sequences for open reading frames, genes, and other features. To these core ACeDB features, LBNL has contributed a versatile query-by-example facility, a mechanism for providing on-line descriptions of data fields, and the new multi-level, multi-map display module.

*This work was supported by the U.S. Department of Energy under Contract Number DE-AC03-76SF00098

Transforming Molecular Biology Databases Using Morphase [1]

S.B. Davidson, A. Kosky, C. Overton and P. Buneman
Dept. of Computer and Information Science & Dept. of Genetics
University of Pennsylvania, Philadelphia, PA 19104
Email: {*susan,coverton,peter*}*@cis.upenn.edu*

The Human Genome Project (HGP) involves a proliferation of different databases, including both archival (GenBank, GSDB) and local notebook databases. These databases frequently use incompatible structures to represent the same or overlapping data, and further may be implemented in a variety of data-models and database management systems (DBMSs), including object-oriented systems (ACEDB, OPM, Object Store), flat-relational databases (SYBASE), and structured text files (ASN.1). It is frequently necessary to transform data between these incompatible databases and data-models. For example data stored in a database at one HGP site may have an impact on the experiments being carried out at another site, and therefore needs to be stored in the local laboratory notebook database at the second site. Further useful tools, such as data browsing or analysis tools, may be implemented for a particular DBMS or database schema, and it is desirable to move data from another database into this system so as to apply these tools. What is required is more than a uniform user interface to distributed databases: data must be integrated and transformed into the structures required by other databases or applications. The problem is aggravated by the rapid evolution of database schemas that results from constantly changing experimental and analysis techniques. Any data transformation system needs to undergo frequent change to reflect these schema evolutions. Finally, there are an increasing number of instantiated integrated genomic databases such as the Integrated Genomic Database (IGD [2]) which require transformations from the member databases into the new structure.

Implementing such transformations by hand on a case by case basis is time consuming and error prone. Consequently there is a need for a method of specifying and implementing transformations in a uniform way, allowing transformations to be specified across a wide variety of different data-models, and to be formally analyzed and verified. Morphase is a prototype system for specifying transformations between data sources and targets in an intuitively appealing, declarative language based on Horn clause logic. Transformation specifications are then translated into an underlying database programming language, CPL[3], for implementation. The data-types underlying Morphase include arbitrarily nested records, sets, variants, lists and object identity, thus capturing the types common to most data formats, in particular ASN.1[4] and ACE[5].

The CPL implementation of Morphase can be connected to a wide variety of data sources through data drivers, modular interfaces that mediate between the internal language of CPL and distributed data sources. Additional drivers for new data sources can easily be added as they arise. In particular, drivers to connect CPL to ASN.1, ACEDB and SYBASE have been developed; other drivers, for example to OPM, are currently being developed. The drivers are used to query as well as update data sources which are instances of their type, e.g. the Sybase driver can be used for our local Sybase database, Chr22DB. In this way, data can be read from multiple heterogeneous data sources, transformed using Morphase according to the desired output format, and inserted into the target data source.

We have tested Morphase by applying it to a variety of different transformation problems involving Sybase, ACE and ASN.1. In particular, we used it to specify a transformation between the Sanger Center's Chromosome 22 ACE database (ACE22DB) and the Philadelphia Genome Center's Chromosome 22 Sybase database (Chr22DB), as well as between a portion of GDB and Chr22DB. Some of these transformations had already been hand-coded without our tools, forming a basis for comparison. Once the semantic correspondences between objects in the various databases were understood, writing the transformation program in Morphase was easy, even by a non-expert of the system. Furthermore, it was easy to find conceptual errors in the transformation specification. In contrast, the hand-coded programs were obtuse, difficult to understand, and even more difficult to debug.

[1] This research was supported by a grant from the Director, Health Effects and Life Science Research Division, Office of Health and Environmental Research of the U.S. Department of Energy under contract DOE DE-FG02-94-ER-61923 Sub 1.

[2] O. Ritter et al., *Computers and Biomedical Research*, 27:97–115 (1994).

[3] P. Buneman et al., *Proceedings of the 21'st International Conference on Very Large Data Bases* (September 1995).

[4] "NCBI ASN.1 Specification", National Library of Medicine, Bethesda, MD (1992).

[5] J. Thierry-Mieg and R. Durbin, "Syntactic Definitions for the ACEDB Data Base Manager" Tech Report MRC Laboratory for Molecular Biology, Cambridge,CB2 2QH, UK, (1992).

Providing Database Access to Structured Files [1]

W. Fan, P. Buneman, S.B. Davidson and C. Overton
Dept. of Computer and Information Science & Dept. of Genetics
University of Pennsylvania, Philadelphia, PA 19104
Email: {peter,susan,coverton,wenfei}@cis.upenn.edu

Much data of interest to biologists exists as text in flat files rather than in database management systems (DBMSs). The reasons for this are numerous, ranging from economic considerations (the expense of a general purpose DBMS), to the numerous special purpose and often quite sophisticated applications that have been built around data stored in these formats (e.g. ACE), to the ubiquity of the format as an exchange format (e.g. ASN.1). Owners of such flat file data sources are therefore frequently reluctant to migrate their data to DBMSs and give up the flat file format. On the other hand, DBMSs offer many desirable features that are not found in general text systems, such as indexing, integrity checking concurrency control and general purpose query languages against which optimizations can be performed. To take advantage of DBMS features while maintaining data in its original format, one solution is to map the file format into a database format and provide a database view for the data in files. In this way, the data can be queried and updated using database languages, and integrity constraints on the data can be maintained using database facilities.

We have therefore developed a framework that, using an extension of Definite Clause Grammars (DCG), translates data stored in text files structured by a syntactic grammar into a database, and converts data from the database back to the files. Our framework is more general than existing mapping mechanisms between files and databases[2,3] in the sense that it allows the grammars specifying files to be beyond context free grammars (CFGs) – in fact, DCG grammars have Turing machine power. Our approach also facilitates general integrity constraints checking on the elements recognized while parsing files. Such features are important since files often cannot be described by CFGs and they encode many integrity constraints. As an example, GenBank cannot be described by a CFG. Furthermore, many constraints are encoded in GenBank accession numbers: they function as keys for entries, and must therefore be unique; they also function as references to other entries when used in non-key fields within an entries (foreign keys). The family of constraints that can be encoded using DCGs is quite general, including but not limited to those commonly found in database systems.

The use of DCGs within our framework also promise new optimization techniques. When mapping files to databases, the size of databases created is a major efficiency concern; ideally, it should be possible to generate the database image for only the data in which the users are interested. Using DCGs, this can be done by describing conditions under which the parser can ignore portions of the input parse, only generating a database image for the data satisfying the conditions. The data which does not satisfy the conditions can simply be skipped in a do-not- care manner. For instance, a GenBank file can consist of several hundred thousand entries. The user can choose to parse N entries only, or to create database image for only the entries satisfying certain conditions. This conditional (partial) parsing technique improves parsing performance and enables users to check constraints without generating database objects.

DCGs can be directly implemented in Prolog, do not require building up special parsers, and thus can be rapidly prototyped. We have also been able to develop a simple technique for "reverse" transformations. That is, given a DCG mapping from files to databases which satisfies certain constraints, another DCG which encodes data from the databases back to the files can be produced which is guaranteed to respect the original grammar.

[1] This research was supported by a grant from the Director, Health Effects and Life Science Research Division, Office of Health and Environmental Research of the U.S. Department of Energy under contract DOE DE-FG02-94-ER-61923 Sub 1.

[2] S. Abiteboul et al., "Querying and updating the file", *Proceedings of VLDB'93*, 73–84 (1993).

[3] G.H. Gonnet and F. W. Tompa, "Mind your grammar: a new approach to modeling text", *Proceedings of VLDB'87*, 339–346, Brighton (1987).

A FREELY SHARABLE SOFTWARE COMPONENT TO MANAGE GENOME-APPLICATION DATA*

Steve Rozen, John Lehman, Lincoln Stein, Nathan Goodman,
{steve,jlehman,lstein,nat}@genome.wi.mit.edu, Whitehead Institute for Biomedical Research, One Kendall Square, Cambridge MA 02139.

We are constructing a data-management component tuned to the requirements of genome applications. This component will offer many of the services commonly provided by database management systems (DBMSs), including

- concurrency control and recovery,
- data-definition and query languages, and
- application-program interfaces supporting client/server operation.

We do not seek to provide extremely high transaction rates and sophisticated query optimizers on a flat data model; existing high-end commercial DBMSs provide such facilities. Instead, the core of this genome data manager is designed to

- support the semantic and object-oriented data models that have been widely embraced for representing genome data,
- provide domain-specific built-in types and operations for storing and querying biomolecular sequences,
- provide built-in support for tracking laboratory workflows, and
- admit further extensions for other special-purpose types.

The core data manager can be customized to support a variety of specific data models—for example ACEDB,[1] OPM,[2] or ODMG-9x[3]—and applications—for example to back WWW servers, for laboratory notebook databases, and for organism and community databases. Since the data manager will be highly portable and free of licensing fees, we expect that it will be attractive as a database for distributing copies of organism or community databases.

We will be reporting progress on the core data manger's architecture and interface at `http://www-genome.wi.mit.edu/informatics/cdm.html`, and we solicit comments on its design. We are currently extending the core data manager to provide ACEDB compatibility for schemas, data-transfer files, client/server interface, and—if needed by potential users—function-call application program interface. We also contemplate extending the core data manager to provide other specific data models depending on the interest of potential users.

The core data manager is being constructed on top of transactional libdb (Berkeley UNIX's `dbopen(3)` with concurrency control and recovery added).[4] We are collaborating with libdb's authors to provide a portable, POSIX-compliant implementation of transactions for this library.

* Supported by a grant from the U. S. Department of Energy under contract DE-FG02-95ER62101.

[1] R. Durbin and J. Thierry-Mieg. A *C. elegans* database, 1991. Documentation, code and data available from anonymous `ftp` servers at `ncbi.nlm.nih.gov`.

[2] I.-M. A. Chen and V. M. Markowitz. An overview of the Object Protocol Model (OPM) and the OPM data management tools. *Information Systems*, 20(5):393–418, July 1995.

[3] R. Cattell, T. Atwood, J. Duhl, G. Ferran, M. Loomis, and D. Wade. *The Oject Database Standard: ODMG-93 Release 1.1.* Morgan Kaufmann Publishers, 1994.

[4] M. Seltzer and M. Olson. LIBTP: Portable, modular transactions for UNIX. *USENIX Winter 1992 Technical Conference*, 1992.

Informatics Support for Mapping in Mouse-Human Homology Regions

Sergey Petrov[1], Manesh Shah[1], Loren Hauser[2], Richard Mural[2], and Edward Uberbacher[1]

Computer Science and Mathematics Division[1] and Biology Division[2], Oak Ridge National Laboratory, Oak Ridge, TN 37831-6364 615/574-6134.

The purpose of this project is to develop databases and tools for the ORNL Mouse-Human Mapping Project, including the construction of a mapping database for the project, tools for management and archiving of cDNAs and other probes used in the laboratory, and analysis tools for mapping, inter-specific backcross, and other needs. A re-evaluation of the needs of the Mouse-Human comparative mapping project has lead us to implement an ACeDB based database for the management and display of the diverse mapping information associated with the project. Our choice of ACeDB for this task was based on the fact that the structure of the database and the organization of the raw data are clear and intuitive, and because it is being used as the data management system for a number of genome projects. Our ACeDB implementation has been modeled somewhat from the chromosome 21 ACeDB system at LBL (with some model modification) and is designed to contain genetic and physical mouse map data as well as homologous human chromosome data. The utility of exchanging map information with LLNL (human chromsome 19) and potentially other centers has lead to the implementation of procedures for data export, and import of human mapping data into the ORNL databases.

Some examples of the information found in the current database are:

1.) Many markers, both mouse and human, have been mapped to Mmu7 using one of two large interspecific backcrosses (M. spretus X M. musculus) created in the laboratory of Dr. Eugene Rinchik. The maps derived from these studies can be displayed in ACeDB and detailed information about the various marker can easily be accessed through this interface. In addition other genetic and physical maps in the system, a map of the homologous region of human chromosome 19q for example, can be directly compared to the genetic map of Mmu7 using the MultiMap feature of ACeDB.

2.) The region around the murine p locus which correspond to the Prader-Willi/Angelman Syndrome region of human 15q, has been extensively studied at Oak Ridge over the last four decades. This has led to a wealth of both physical and genetic data which has been incorporated into the current database and which can be viewed and accessed through the ACeDB interface.

3.) A genetic map of Mmu7 based on simple-repeat polymorphisms generated at MIT has been added to the database along with a MultiMap comparing it to the Oak Ridge IB map of Mmu7. This is a step toward integrating data being generated in other parts of the mouse community with the information being generated at Oak Ridge.

User access to the system is being provided by workstation forms-based data entry and ACeDB graphical data browsing. We have also implemented the LLNL databases browser to view the human chromosome 19 data maintained at LLNL, and arrangements are being made to incorporate mouse mapping information into the browser. Other applications such as the "Encyclopedia of the Mouse", specific tools for archiving and tracking cDNAs and other mapping probes, and analysis of inter-specific backcross data and restriction mapping of YACs have been implemented.

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

AUTOMAIL: AN AUTOMATED SYSTEM FOR MANAGING EMAIL QUERIES TO SERVERS*

Elaine Best, Joe Gatewood, Computing, Information and Communications Division and Life Sciences Division, Los Alamos National Laboratory

Automail is a UNIX package which automates the process of sending queries to an email-based server, tracking the messages, and then handling the response. It accepts sequences either as a list of files or from a SYBASE database. The responses are either stored in the SYBASE database or placed in a directory accessible by the user. Once the user submits the sequences, the package runs periodically without human intervention.

Automail offers the following useful features:

- Given a sequence, Automail can format the query in the style required by each server.
- Each query is logged and tracked. If a response is not received within a designated time period, the query is resent.
- Automail avoids flooding the server with a large number of queries at once. No more than a designated number of queries to a server are permitted to be outstanding. As responses are received, new queries are sent out.
- Automail works around the clock, making use of nighttime and weekend periods, when the servers may be less busy.
- If the database option is chosen, Automail will parse the response and store it in a database.

Automail is currently configured to work with Genquest in the file-based option, and cDNA-Inform for the database option. However, new servers can be added with a minimum of trouble.

*This work was funded by USDOE under contract W-7405-ENG-36.

System Design of The Genome Topographer

S. Cozza, D. Cuddihy, R. Iwasaki, M. Mallison, C. Reed, J. Salit, A. Tracy, T. Marr.
Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724.

Genome Topographer (GT) is an advanced genome informatics system that has received joint funding from DOE and NIH over a number of years. DOE funding has focused on GT tools supporting computational genome analysis, principally on sequence analysis. GT is scheduled for public release next spring under the auspices of the Cold Spring Harbor Human Genome Informatics Research Resource. GT has 17 major existing frameworks: 1. Views, including printing, 2. Default manager,3. Graphical User Interface, 4. Query, 5. Project Manager, 6. Workspace Manager, 7. Asynchronous Process Manager, 8. Study Manager, 9. Help, 10. Application, 11. Notification, 12. Security, 13. World Wide Web Interface, 14. NCBI, 15. Reader, 16. Writer, 17. External Database Interface. GT Frameworks are independent sets of VisualWorks (client) or SmallTalkDB (GemStone) classes which interact to perform the duties required to satisfy the responsibilities of the specific framework. Each framework is clearly defined and has a well-defined interface to use it. These frameworks are used over and over in GT to perform similar duties in different places. GT has basic tools and special tools. Basic tools get used many times in different applications, while special tools tend to be special purpose, designed to do fairly limited things, although the distinction is somewhat arbitrary. Tools typically use several frameworks when they get assembled. Basic Tools: 1. Project Browser, 2. Editor/Viewer, 3. Query, 4. NCBI Entrez, 5. File reader/writer, 6. Map comparison, 7. Database Administrator, 8. Login, 9. Default, 10. Help. Special Tools: 1. Study Manager, 2. Compute Server, 3. Sequence Analysis, 4. Genetic Analysis. These frameworks and tools are combined with a comprehensive database schema of very rich biological expression linked with plugable computational tools. Taken together, these features allow users to construct, with relative ease, on-line databases of the primary data needed to study a genetic disease (or genes and phenotypes in general) from the stage of family collection and diagnostic ascertainment through cloning and functional analysis of candidate genes, including mutational analysis, expression information, and screening for biochemical interactions with candidate molecules.GT was designed on the premise that a highly informative, visual presentation of comprehensive data to a knowledgeable user is essential to their understanding. The advanced software engineering techniques that are promoted by using relatively new object oriented products has allowed GT to become a highly interactive and visually-oriented system that allows the user to concentrate on the problem rather than on the computer. Using the rich data representational features characteristic of this technology, the GT software enables users to construct models of real-world, complex biological phenomena. These unique features of GT are key to the thesis that such a system will allow users to discover otherwise intractable networks of interactions exhibited by complex genetic diseases.

The VisualWorks development environment allows the development of code that runs unchanged across all major workstation and personal computers, including PCs, Macintoshes and most Unix workstations. Supported by grants DE-FG02-91ER61190 and P41 HG01268-01

Layout Tools for Clone End-Sequence Sampling Contigs*

Guochun Xie, Michael L. Engle, and Christian Burks

Theoretical Biology and Biophysics Group & Center for Human Genome Studies; T-10, MS K710; Los Alamos National Laboratory, Los Alamos, NM 87545.

Several groups have developed variations on the general strategy of clone end-sequence sampling [1-4], with the goal -- given a target cloned region -- of using the assembly of end-sequences of sub-clones, information about sub-clone lengths, and offset and orientation relationships among pairs of end-sequences to drive a sub-clone layout. Most sequence assembly packages currently available do not provide for this 'meta-assembly' task. Generating layouts for and among the sequence contigs in this context is part of the general challenge of taking advantage of ancillary information during sequence assembly [5]. Using the cosmid-based SASE (SAmple SEquencing) data sets at Los Alamos [4] as a starting point, we have developed a simple suite of tools for: (i) extracting the end-sequence contig information from ABI Autoassembler data files; (ii) linking related end-sequence contigs to one another based on sub-clone assignments; (iii) developing a layout for the contigs based on the links established in (ii) and sub-clone lengths; and (iv) display of the layouts developed in step (iii) in an X-Windows plot. These tools are meant for use during in-stream evaluation of the coverage and experimental consistency of sampled sequence data. The generic specification of objects in the input file for the layout module in (iv) should make it useful for graphically displaying layout relationships among a variety of different types of sequencing and mapping data. We will describe these tools in greater detail and show examples of their application to clone end-sequencing sample data.

*This work was done under the auspices of the Department of Energy, and was supported through the DOE/OTHER genome project (R. Moyzis, P.I., ERW-F137).

[1] Chen EY; Schlessinger D; and Kere J. (1993) Ordered shotgun sequencing, a strategy for integrated mapping and sequencing of YAC clones. *Genomics* **17**: 651-656.

[2] Smith MW; Holmsen AL; Wei YH; Peterson M; Evans GA. (1994) Genomic sequence sampling: a strategy for high resolution sequence-based physical mapping of complex genomes. *Nature Genetics* **7**: 40-47.

[3] Roach JC; Boysen C; Wang K; and Hood L. (1995) Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics* **26**: 345-353.

[4] Moyzis, RK; Doggett, NA; Altherr, MR; and Deaven, LL. (1995) An integrated physical map of human chromosome 16: Sample Sequencing (SASE) analysis as a framework for complete genomics sequencing. *Genome Science & Technology* **1**: P-18.

[5] Burks C; Parsons, RJ; and Engle, ML. (1994) Integration of competing ancillary assertions in genome assembly. In "*Proceedings: Second International Conference on Intelligent Systems for Molecular Biology*", Altman et al., Eds. AAAI Press, Menlo Park, CA, pp. 62-69.

Software Support for High-Throughput Sequence Reconstruction*

Charles Lawrence, Victor Solovyev, Ellen Bergeman, and Pam Culpepper, Department of Cell Biology, Baylor College of Medicine, Houston, TX 77030.

The Genome Reconstruction Manager (GRM) is an advanced software component for managing the data flow for high-throughput sequencing. A new release is currently available that includes the following features:

- Automation of preprocessing and assembly data flow.
- Computation of position-specific error probabilities for ABI sequence data.
- Use of position-specific error probabilities to support semi-automated editing of contigs.
- Use of position-specific error probabilities to support probability-based consensus computation and reporting of consensus sequence with confidence values.
- Computation of 'super contig' structure from pairwise distance and orientation constraints.
- Sequence assembly using the FAK 3.0 assembly kernel developed by E. Myeres and coworkers.
- Client-server based assembly (assembly may be performed on networked Sun or DEC Alpha hosts).
- Flexible visualization and interactive editing of contig multialignments and consensus sequence with coordinated display of aligned fluorescent traces.
- Seamless integration of all steps in a single application.

GRM runs on Sun workstations running Solaris 2.3 or 2.4.

For more information look at URL *http://gc.bcm.tmc.edu:8088/GRM/grm-home.html* or contact C. Lawrence (*chas@bcm.tmc.edu*).

*Supported by a grant from the DOE Human Genome Program, DE-FG-94ER61818

Lawrence, C.B, Honda, S., Parrott, N.W., Flood, T.C., Gu, L., Zhang, L., Jain, M., Larson, S., and Myers, E.W. 1994. The Genome Reconstruction Manager: A software environment for supporting high-throughput DNA sequencing. *Genomics* 23:192-201.

Lawrence, C.B. and Solovyev, V.V. 1994. Assigning position-specific error probabilities to primary DNA sequence data. *Nucl. Acids Res.* 22:1272-1280.

A Unix-based Graphical Editing Tool for Shotgun DNA Sequence Assemblies

Chris Abajian, David Gordon, Leroy Hood, Phil Green. Department of Molecular Biotechnology, University of Washington, Seattle WA

Two approaches to reducing the cost of finishing shotgun sequence assemblies are 1) improving the quality of the initial computer-generated assembly, reducing the amount of human intervention required and 2) optimizing the user interface to maximize efficiency of the finisher's effort. Phil Green has written an assembly program, Phrap, that performs very well compared to other commonly-used assembly programs but does not provide for editing or graphical display of the assembly. Consed ("CONSensus EDitor") is a graphical editing tool that allows users to view and edit the output of Phrap. It takes advantage of Phrap's assembly algorithm and provides an editing interface designed in collaboration with experienced finishers in the department. Consed is written in C++ and is supported on a variety of platforms, including Solaris, SunOs, HP-UX and OSF-1.

COMBINATORIAL ASPECTS OF MULTIPLE-COMPLETE-DIGEST RESTRICTION MAPPING

Richard M. Karp, Department of Computer Science and Engineering, University of Washington and Geoffrey Zweig, International Computer Science Institute, Berkeley, CA[1]

Restriction mapping is the process of determining the restriction sites on a target DNA molecule associated with one or more restriction enzymes. Maynard Olson is leading a project at the University of Washington in which restriction maps for two or more restriction enzymes are constructed by the following process:

1. Construct a clone library giving 15-20X coverage of the target molecule.
2. Completely digest each clone with each restriction enzyme and measure the fragment sizes using gel electrophoresis.
3. From this data, reconstruct the overlap structure of the clones and the positions of the restriction sites for each restriction enzyme.

Our research is concerned with two computational aspects of this multiple-complete-digest restriction mapping problem: determining clone overlaps and fragment identification.

Detecting Clone Overlaps Given a large clone library, the goal is to determine those pairs of clones that are highly likely to overlap. We would like to discover these likely overlaps without explicitly comparing each pair of clones, since such a comparison process would be extremely time-consuming when the number of clones is very large.We have devised a randomized algorithm which finds the highly overlapping pairs without resorting to exhaustive comparisons [2]

Fragment Identification Once the overlaps among the clones have been constructed various greedy algorithms can be used to determine the ordering of the clones along the DNA. More difficult, however, is the construction of the restriction map given the clone ordering. The central problem is fragment identification - partitioning the fragment occurrences on the individual clones into "equivalence classes," each of which corresponds to a single physical restriction fragment. We have characterized the conditions under which a partitioning of the fragment occurrences is realizable as a physical map. We are currently implementing fragment identification algorithms based on this characterization.

[1]Research supported by DOE Grant 03-94ER61913.000

[2]R.M. Karp, O. Waarts and G. Zweig,*Proc. IEEE Symp. on Foundations of Computer Science* 621-630 (1995)

Hopper: A Prototype for Data Flow in Large-Scale DNA Sequencing

Todd M. Smith[1], *Chris Abajian, Leroy Hood*, Department of Molecular Biotechnology, University of Washington, Seattle WA, 98195

With the advent of fluorescent-based DNA sequencing it has become possible to consider the analysis of entire genomes as a first step in the biological study of an organism. A primary challenge in these projects is the scale at which the data handling must be done. For example, at least 10^7 sequencing tracts will have to be obtained to completely determine the nucleotide sequence of the human genome. Hence, large-scale sequencing facilities will benefit from tracking template DNA information (purification methods, reaction, and electrophoresis conditions), in a systematic fashion. A lack of software tools that support automated sample entry however is a major hindrance to recording these parameters. For example, experimental information can be added to the comment field in the ABI sample sheet, and subsequently the chromatogram file, but this information must be added by hand. We have overcome this problem, with a sample sheet generator that uses the ABI file format, written in a graphical programming language, Tcl/Tk. It is used to facilitate data flow in our production operation.

The UNIX file system has been used to prototype automating the flow of data from the ABI sequencer to a data repository. Data transfers between an Apple Macintosh (the collection device for the ABI sequencer) and a UNIX workstation are accomplished by FTP (file transfer protocol) using a Macintosh program, Fetch. Once transferred, the data are automatically processed by a central Perl program, Hopper. Hopper automatically runs a series of programs that provide a number of first level analysis about data quality (read length estimate, fraction of indeterminate bases, and number of contaminating and repetitive sequences) and generates simple reports describing the results. This program also automates DNA sequence data assembly using the PHRED[2] basecalling and PHRAP[2] assembly programs. Using the combination of PHRED and PHRAP cosmids, from shotgun sequencing projects (containing up to 40% alu repetitive DNA), have been successfully assembled without manual intervention, as well as BAC derived contigs over 100 kb in length.

Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy under contract

[1] DOE Human Genome Distinguished Postdoctoral Fellow.

[2] Phil Green unpublished work.

QVIEW: A SOFTWARE TOOL FOR THE QUANTITATIVE ANALYSIS OF PROTOCOL SUCCESS, BASE-CALLING ALGORITHMS AND QUALITY CONTROL FOR LARGE-SCALE SEQUENCING PROJECTS*

James N. Labrenz[1] *and Tim Hunkapiller,* University of Washington, Department of Molecular Biotechnology, Box 357730, Seattle, WA 98195.

As our technical ability to generate DNA sequence data increases at an ever expanding rate, debate persists as to the best choice of strategies to pursue in order to provide both optimal biology and economy within genome-scale projects. Although many technical issues remain to be resolved, it is first critical to define the "end product" that minimally satisfies the community's objective. In other words, in order to determine the best strategy, we need first to define the model of completeness and accuracy we require of the final or representational sequence for a given portion of the genome. The debate can be intense here as well, for there are many opinions as to just how important accuracy of the final sequence is in defining its biological 'reality' or at least its usefulness.

Unfortunately, because of the significant subjective component of data analysis of traditional sequencing methods, these examinations are difficult to pursue as so little is understood of the true nature of error in the raw data, let alone its relationship to the accuracy of the final sequence. We are pursuing a rigorous examination of principal DNA sequence data of a large-scale, genomic sequencing project generated with automated sequencing instruments under various protocol regimens. Our objective is to establish the nature of variation between the raw and consensus data in order to (1) establish better rules for translating the raw instrument data into called bases; (2) provide a quantitative basis for developing 'confidence' values for raw base calls; (3) compensate for the impact of error on feature-identification (gene finding, database comparisons, etc.); and (4) provide a suite of software and database tools that will provide for the consistent and automated error analysis of large amounts of DNA sequence data. This tool will allow researchers to better evaluate variation in laboratory procedures (which polymerase is best, are long gels worth it, etc.), to evaluate the efficacy of different base calling methods and to assess the efficiency of various algorithms critical to successful sequence assembly (end clipping, repeat sequence discrimination, overlap analysis, etc.). In addition, the tool will provide a quality control evaluation for the day-to-day sequencing effort.

It is expected that these efforts will provide insight into the biology of sequencing (i.e., how polymerases interact with their substrate), help in the development of better tools for data characterization and provide a quantitative approach to protocol optimization. It is assumed that a better understanding of the raw DNA sequence data, will allow for the extraction of more useful information with the same amount of laboratory effort, hence impacting on the choice of sequencing strategies and the economies of large-scale projects. We report here on our software development efforts as well as representative analyses of data from a mega-base sequencing project.

*Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy under contract DE-

[1]DOE Human Genome Distinguished Postdoctoral Fellow.

AN INFORMATION MANAGEMENT SYSTEM FOR DNA SEQUENCING

Arthur Kobayashi (kobayashi1@llnl.gov), David J. Ow, T. Mimi Yeh, Mark C. Wagner, Thomas R. Slezak. Human Genome Center, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94550.

We are implementing a comprehensive system to track and manage DNA sequencing data. The major goals of this system are to manage a complete source, processing, and analysis history for DNA sequenced in our local laboratory, and to streamline the flow of information and products through the laboratory.

To accomplish these goals we have developed tools to define and build clone library hierarchies to track detailed information at each processing and sequencing step, and other tools to help setup and archive results from our laboratory instruments. All information is stored in a relational database (Sybase). We have designed the system to be compatible with existing laboratory functions and protocols, minimize data entry, provide for consistent naming conventions, automate generation of sample sheets and setup files, and to track replications (reprocessing) of items.

There are numerous components that make up this system. The main interface, which runs on a Unix workstation, is used to create and maintain processing and sequencing information. Macintosh-based programs edit labelling or sequencing runs, update the database, and create final sample sheets or configuration files used to set up labellers and sequencers. We also use custom forms on Web browsers (such as Mosaic or Netscape) to implement specialized functions such as creating clone library entries or editing certain types of sequencing run assignments. We have developed other tools to streamline ABI setups and data archival using various Macintosh scripting tools.

We have been implementing this system using a variety of methods, including C, X Windows/Motif, UIMX (a commercial graphical-user interface builder), custom Excel spreadsheets, Sybperl, MacPerl, HTML (Hypertext Markup Language), and AppleScript. The resulting system is distributed, heterogeneous, and very loosely-coupled. Functions may be implemented as stand-alone, single-purpose programs, or as larger, multi-function programs; all of the pieces of the system are integrated over our computer network and share the use of an underlying common schema and relational database. This approach gives us great flexibility in selecting the most effective means available to implement a given function, and to replace or modify portions of the system without impacting other parts of the system.

The system is now in routine use in our laboratory. We are continuing to refine and improve existing functions, and are currently adding capabilities to streamline and integrate the analysis of sequenced DNA.

(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.)

TOWARDS AUTOMATIC ASSEMBLY FOR THE DIRECTED SEQUENCING STRATEGY*

Samuel Pitluck, Arun Aggarwal, Frank Eeckman, Eugene Veklerov, Human Genome Informatics Group, Lawrence Berkeley National Laboratory, Berkeley CA 94720

The Lawrence Berkeley National Laboratory Human Genome Center pioneered the directed sequencing strategy. Using this strategy, we map all the sequence initiation sites using transposons. This information can be used to facilitate the assembly process. We decided to use the Fragment Assembly Kernel (FAK-Larson,Jain,and Meyers, 1994) in our assembly procedures because it can handle up-front mapping information as constraints. FAK has been incorporated into two C-language programs. One program (SPASS) is used to assemble 300-450 bp fragments into contigs of 3000-4500 bp in length. The second program is used to find a tiling path for our double ended sequencing strategy.

Our assembly program is able to make use of the information that is available about the fragments that come from transposon pairs. Each transposon site yields two reverse complimented sequences that overlap by 5 bp. The information for each transposon site is summarized in a constraint file. This constraint file along with all sequence fragments is read into our assembly program for processing. We used SPASS to assemble the 3 kb subclones and compared its performance to XBAP (Staden, 1992). We report on the results of this comparison. In general SPASS has produced fewer contigs than XBAP.

We also built an interface to FAK using SPACE, a variant of ACeDB (Durbin and Mieg, 1991). ACeDB has been used mostly as a database program. In SPACE we have added the capability of trace editing, assembly, as well as fragment and contig display. Because FAK consists of a library of functions it is easy to customize both the input and output to and from the assembler. The assembler package communicates with SPACE via constraint and fragment files written in .ace format. This process is transparent to the user. Thus, users are able to select fragments for assembly and call on our assembler to assemble the fragments. The results can then be readily displayed within SPACE.

We have also developed an algorithm to find tiling paths used in our double ended sequencing strategy. Here, we used the compare function in FAK to find "hits" between all the end fragments of all 3 kb subclones in a particular 80 kb P1 clone. We use 192 random clones and generate 384 end fragments from these. The compare function returns a score for each comparison. The score "roughly reflects the length of the overlap with a deduction for mismatches in the alignment." We accept two fragments as overlapping if the score is greater than 15. After all the "hits" are determined, a separate program extracts all the possible tiling paths. We present examples of these comparisons and tiling paths.

* This work was supported by the U.S. Department of Energy under Contract Number DE-AC03-76SF00098

A Workbench for Sequence Annotation and Browsing

Nomi L. Harris and Frank H. Eeckman, Human Genome Informatics Group*, Lawrence Berkeley National Laboratory, MS 46A/1123, Berkeley CA 94720; nlharris@lbl.gov

Sequencing centers such as the Human Genome Center at LBNL are producing an ever-increasing flood of genetic data. Annotation can greatly enhance the biological value of these sequences. Useful annotations include homologies to known genes, possible gene locations, gene signals such as promoters, etc.

We are developing a workbench for automatic sequence annotation and annotation viewing and editing. The goal is to run all available sequence analysis tools and display the results in such a way that the various predictions can be compared. Researchers will then be able to examine all of the annotations (for example, the genes predicted by various gene-finding methods) and select the ones that look the best.

Our current prototype annotation workbench automatically runs the following sequence analysis tools:

HOMOLOGY SEARCHES
- blastx (against all available sequences translated to amino acids)
- blastn (against the EST database)
- blastn (against human repeat sequences)
- trnascan (to look for tRNA genes)

GENE FINDING
- Genefinder (with human tables)
- GRAIL
- xpound

PROMOTER PREDICTION
- neural net promoter prediction

The resulting predictions are filtered and saved in simple data formats such as .ace format. Other sequences analysis tools can also be incorporated.

The choice of sequence analysis programs is orthogonal to the front end used to view them. We have developed a prototype annotation browser based on the bioTkperl map display widget written by David Searls and Gregg Helt. Color-coded sequence annotations for both strands are displayed on a canvas that can be scrolled and zoomed. Clicking on an annotation displays additional information about it.

Planned extensions to ACEDB will enable it to serve as an alternative annotation browser.

In order to test our sequence analysis environment, we used it to study the HUM14SP6 region of human chromosome 5 (5q31), which was sequenced at LBNL. (In another abstract[1] at this meeting, the biology group will present their findings on a larger section of 5q31.) Although the 22Kb HUM14SP6 segment belongs to a region that contains many interleukin genes, no genes have yet been identified in HUM14SP6.

HUM14SP6 was found (by BLAST) to have 552 hits (significant regions of homology) with ESTs and 445 hits with sequences in a non-redundant amino acid database (NRDB), many of which were similar or identical to the EST hits. 47 homologies with human repeat sequences covered roughly half of the EST/NRDB hits.

The three gene-finding programs we ran each found several possible exons in the complementary strand (Genefinder and GRAIL also found one possible exon in the forward strand). Half of GRAIL's 12 predicted exons overlapped with blast hits, as did all but two of Genefinder's 10 predicted exons. Most of xpound's predictions echoed GRAIL's. The promoter predictor found numerous possible promoters in both strands.

---

* This work was supported by the U.S. Department of Energy under Contract Number DE-AC03-76SF00098

[1] Kelly A. Frazer, Yukihiko Uedo, Maria R. Garofalo, Jan-Fan Cheng, and Edward M. Rubin, "Computational and biological analysis of 1.2 Mb of sequence at 5q31", this meeting.

A CUSTOMIZABLE SOFTWARE SYSTEM FOR FRAGMENT ASSEMBLY†

Gene Myers and Susan Larson, Department of Computer Science, University of Arizona, Tucson, AZ 85721

We have completed the design and begun construction of a software environment in support of DNA sequencing called the "FAKtory". The environment consists of (1) our previously described software library, FAK, for the core combinatorial problem of assembling fragments, (2) a Tcl/Tk based interface, and (3) a software suite supporting a modest database of fragments and a processing pipeline that includes clipping and vector prescreening modules. A key feature of our system is that it is highly customizable: the structure of the fragment database, the processing pipeline, and the operation of each phase of the pipeline are specifiable by the user. Such customization need only be established once at a given location, subsequently users see a relatively simple system tailored to their needs. Indeed one may direct the system to input a raw dataset of say ABI trace files, pass them through a customized pipeline, and view the resulting assembly with two button clicks.

The system is built on top of our FAK software library and as a consequence one receives (a) high-sensitivity overlap detection, (b) correct resolution to large high-fidelity repeats, (c) near perfect multi-alignments, and (d) support of constraints that must be satisfied by the resulting assemblies. The FAKtory assumes a processing pipeline for fragments that consists of an `INPUT` phase, any number and sequence of `CLIP`, `PRESCREEN`, and `TAG` phases, followed by an `OVERLAP` and then an `ASSEMBLY` phase. The sequence of clip, prescreen, and tag phases is customizable and every phase is controlled by a panel of user-settable preferences each of which permits setting the phase's mode to `AUTO`, `SUPERVISED`, or `MANUAL`. This setting determines the level of interaction required by the user when the phase is run, ranging from none to hands-on. Any diagnostic situations detected during pipeline processing are organized into a log that permits one to confirm, correct, or undo decisions that might have been made automatically.

The customized fragment database contains fields whose type may be chosen from `TIME`, `TEXT`, `NUMBER`, and `WAVEFORM`. One can associate default values for fields unspecified on input and specify a control vocabulary limiting the range of acceptable values for a given field (e.g., `John`, `Joe`, or `Mary` for the field `Technician`, and `[1,36]` for the field `Lane`). This database may be queried with SQL-like predicates that further permit approximate matching over text fields. Common queries and/or sets of fragments selected by them may be named and referred to later by said name. The pipeline status of a fragment may be part of a query.

The system permits one to maintain a collection of alternative assemblies, to compare them to see how they are different, and directly manipulate assemblies in a fashion consistent with sequence overlaps. The system can be customized so that *a priori* constraints reflecting a given sequencing protocol (e.g. double-barreled or transposon-mapped) are automatically produced according to the syntax of the names of fragments (e.g. `X.f` and `X.r` for any `X` are mates for double-barreled sequencing). The system presents visualizations of the constraints applied to an assembly, and one may experiment with an assembly by adding and/or removing constraints. Finally, one may edit the mutli-alignment of an assembly while consulting the raw waveforms. Special attention was given to optimizing the ergonomics of this time-intensive task.

† Supported by DOE grant DE-FG03-94ER61911.

Towards completely automated sequence assembly

Phil Green, Molecular Biotechnology Department, Univ. of Washington, Seattle.

The human genome project is moving into its decisive final phase, in which the genome sequence will be determined in large-scale efforts carried out in a number of laboratories. Although current technology appears largely adequate to the task, it will be essential to reduce as much as possible the need for skilled human labor. Editing (correction of base calls and assembly errors) is at present one of the most skill-intensive aspects of genome sequencing, and as such is a bottleneck to increased throughput, a potential source of uneven sequence quality, and an obstacle to more widespread participation in genomic sequencing by the community. We are working towards the long term goal of completely removing the need for human intervention at this stage, with the short-term goals of improving the accuracy of assembly and base-calling, and of more precisely delineating sequence regions requiring human review.

We have developed a program (phred) for making improved base calls and quality assessment of processed ABI 373A and 377 trace data, and an assembly program (phrap). Overall, phred base calls have approximately 40% fewer errors than ABI base calls. Phred's quality measures, which take into account peak spacing, the location of unresolved peaks, and the size of any uncalled peaks, allow identification of subsets of the read having error rates of specified levels. In typical data sets, about 25% of the usable read length consists of bases that can be identified as having an error rate less than 1 per 10kb, and about 60% is identifiable as having an error rate less than 1 per kb. This has important implications for the depth of coverage required in shotgun sequencing, since it implies that low error rates may be attainable when some regions are single-stranded.

Phrap uses quality information, both direct (from phred analysis) and indirect (from read comparison), to delineate the likely accurate base calls; this helps distinguish repeats, and permits use of the full (untrimmed) reads in assembly. In outline, the key assembly steps are as follows: (1) Reads are compared pairwise using a fast implementation of the Smith-Waterman algorithm. Alignment scores are then adjusted to reflect the qualities of discrepant bases, and the list of matches is ranked by these adjusted scores. At this stage anomalous reads (e.g. chimeras) are also identified. (2) A greedy assembly algorithm is used to construct a layout of read overlaps, based on the pairwise comparisons. (3) The contig sequence is constructed from the layout as a "mosaic" of the highest quality parts of the reads; this is done by finding an optimal path through an appropriately defined weighted directed graph. (4) The quality of the assembly is analyzed by enumerating discrepancies between reads and the contig sequence, "weak joins" that are potential sites of misassembly, and consistency of forward/reverse read pairs. (5) A probability of error (reflecting the amount and quality of trace data) is computed for each sequence position. This can be used to focus human editing on particular regions, and to automate decision-making about where additional data is needed.

In collaboration with L. Rowen and with the St. Louis / Sanger consortium, we have begun systematic studies of the performance of these programs on representative cosmid datasets. For 9 mammalian and 9 *C. elegans* cosmids, the complete (final, but unedited) sets of ABI traces were analyzed using phred to obtain base calls and quality measures, and the reads were reassembled using phrap. In each case, all reads (apart from chimeras and singlet or doublet "contaminants") assembled into 1 or 2 contigs. There were no false joins. The per base error rates (relative to the human edited standard) for the 18 cosmids averaged 1 error per 4 kb, with less than 1 error per 20 kb in the phrap "high quality" bases (which constitute 95% of the total sequence).

These results suggest that it should be possible to substantially reduce editing labor in the near future without significantly compromising sequence accuracy.

TIGR Assembler:
A New Tool for Assembling Large Shotgun Sequencing Projects

Granger G. Sutton, Owen White, Mark D. Adams and Anthony R. Kerlavage, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850

In large shotgun sequencing projects DNA fragments are assembled into a consensus sequence. The basic approach is to compare each pair of fragments to find overlaps and use this information to build a consensus sequence. Two obstacles are the large number of pairwise comparisons and the presence of repetitive elements. TIGR Assembler[1] uses a fast initial comparison of fragments (similar to BLAST) to eliminate the need for a more sensitive comparison between most fragment pairs greatly reducing the computer search time. TIGR Assembler recognizes potential repetitive elements by determining which fragments have more potential overlaps than expected given a random distribution of fragments. Repetitive elements are dealt with in a number of ways: repetitive regions are assembled last so that maximum information from non-repetitive regions can be used, the stringency of match criteria is increased in repetitive regions, and constraints involving fragments sequenced from both ends of a clone are used. Short repetitive elements less than half the length of the average fragment are usually not a problem because they are most often spanned by a single fragment. Likewise, repetitive elements which are significantly less similar than the fragment sequencing accuracy (e.g. 94% similar vs. 98% accurate) can be handled by increasing the match stringency. For long, nearly identical repetitive elements sequencing from both ends of clones of known average length and reasonably small variance is essential. This allows fragments which are totally contained in a repetitive element to be properly placed by TIGR Assembler based on the position of their corresponding clone mate. This technique will not work for repetitive regions longer than the average clone length. For very long, nearly identical repetitive regions a second library of much longer clones sequenced from both ends is necessary for TIGR Assembler to determine which flanking regions should be joined. TIGR Assembler can fill the very long repetitive regions with a consensus sequence or the exact sequence can be determined by walking the repeat containing clone. The basic steps in the TIGR Assembler algorithm are as follows: 1) perform pairwise fragment comparisons for the entire data set to generate a list of potential fragment overlaps. 2) use the distribution of the number of potential overlaps for each fragment to label fragments as repeat or non-repeat. 3) start with a non-repeat fragment as the initial assembly seed or a repeat fragment if no non-repeat fragment is left; quit if no fragments remain. 4) use potential overlap list to attempt merges between the current assembly and non-repeat fragments. 5) when no potential overlaps with non-repeat fragments remain for the current assembly, increase the stringency of the match criteria and enforce clone length constraints when attempting to merge with repeat fragments. 6) if due to a merge with a repeat fragment, a non-repeat fragment is added to the potential overlap list go to step 4. 7) when there are no fragments left on the current potential overlap list, output information about the current assembly and go to step 3. TIGR Assembler has been used to assemble the complete genomes of *H. influenzae* and *M. genitalium*.

[1]Sutton G., White O., Adams M. and Kerlavage A., (1995), TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects, *Genome Science & Technology*, 1(1), 9-19.

Analysis and Annotation of Nucleic Acid Sequence

David J. States, Ron Cytron, Pankaj Agarwal and Hugh Chou
Institute for Biomedical Computing, Washington University in St. Louis

URL: http://ibc.wustl.edu email: states@ibc.wustl.edu

Bayesian estimates for sequence similarity: There is an inherent relationship between the process of pairwise sequence alignment and the estimation of evolutionary distance. This relationship is explored and made explicit. Assuming an evolutionary model and given a specific pattern of observed base mismatches, the relative probabilities of evolution at each evolutionary distance are computed using a Bayesian framework. The mean or the median of this probability distribution provides a robust estimate of the central value. Bayesian estimates of the evolutionary distance incorporate arbitrary prior information about variable mutation rates both over time and along sequence position, thus requiring only a weak form of the molecular-clock hypothesis.

The endpoints of the similarity between genomic DNA sequences are often ambiguous. The probability of evolution at each evolutionary distance can be estimated over the entire set of alignments by choosing the best alignment at each distance and the corresponding probability of duplication at that evolutionary distance. A central value of this distribution provides a robust evolutionary distance estimate. We provide an efficient algorithm for computing the parametric alignment, considering evolutionary distance as the only parameter.

These techniques and estimates are used to infer the duplication history of the genomic sequence in C. elegans and in S. cerevisae. Our results indicate that repeats discovered using a single scoring matrix show a considerable bias in subsequent evolutionary distance estimates.

Model based sequence scoring metrics: PAM based DNA comparison metric has been extended to incorporate biases in nucleotide composition and mutation rates, extending earlier work (States, Gish and Altschul, 1993). A codon based scoring system has been developed that incorporates the effects biased codon utilization frequencies.

A dynamic programming algorithm has been developed that will optimally align sequences using a choice of comparison measures (non-coding vs. coding, etc.). We are in the process of evaluating this approach as a means for identifying likely coding regions in cDNA sequences.

Efficient sequence similarity search tools: Most sequence search tools have been designed for use with protein sequence queries a few hundred residues long. The analysis of genomic DNA sequence necessitates the use of queries hundreds of kilobases or even megabases in length. A memory and computationally efficient search tool has been developed for the identification of repeats and sequence similarity in very large segments of nucleic acid sequence. The tool implements optimal encoding of the word table, repeat filters, flexible scoring systems, and analytically parametrized search sensitivity. Output formats are designed for the presentation of genomic sequence searches.

Federated databases: A sybase server and mirror for GSDB are being developed to facilitate the annotation of repeat sequence elements in public data repositories.

Statistical Interpretation of Aligned Sequences Related by Evolution

William J. Bruno, Theoretical Biology and Biophysics (T-10), MS K-710
Los Alamos National Laboratory, Los Alamos, NM 87545; billb@LANL.GOV.

A great deal of information is in principle contained in the evolutionary history of a gene or DNA regulatory element. Such a history may be viewed as a vast series of point mutation experiments, with successful variants being retained and unsuccessful ones being removed from the ensemble.

A simple quantity to investigate is the frequency of each nucleotide in each position of an alignment, corrected for sample bias caused by the evolutionary interrelationships of the sequences. These corrected frequencies should correspond to the frequencies one would observe in a large set of very distantly related sequences sharing a common function, and they may be interpreted as the "fitness" of a nucleotide in a given position. Previous methods for estimating such a fitness have relied on heuristic "sequence weighting" methods to correct for evolutionary relationships.

A more general, likelihood-based approach to estimating corrected nucleotide frequencies is presented. The method employs a modified EM algorithm to generate estimates of the corrected frequencies, as well as an estimate of the phylogenetic tree relating the sequences. Tree topologies are supplied to the program, either by the user or automatically by calls to existing distance-based phylogeny programs, and branch lengths are optimized taking the nucleotide frequencies at each position into account. Although the underlying model of mutation and selection is highly simplified, it captures the discrete nature of the process.

The resulting model for the constraints on the sequences can be used to improve their alignment. Furthermore, the amount of covariation between different sites in a sequence—corrected for evolutionary relationships—can be estimated.

This work supported in part by DOE contract W-7405-ENG-36.

GENERALIZED HMM'S AND THE ANALYSIS OF DNA SEQUENCES

Catherine A. Macken and Kevin P. Murphy, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545.

Hidden Markov Models (HMMs) have received considerable attention in the recent literature as models underlying recognition of patterns in DNA and RNA, gene finding, speech recognition and other applications. With few exceptions, the approach adopted has required prespecification of the topology of the HMM. Since HMMs have typically been used in structured contexts, such as on a family of protein sequences to discover motifs, or to predict the two-dimensional structure of RNA, there is often a natural topology that can be specified. The analysis then requires inference of the independent probabilities of transition among the hidden states and of symbol emission, that is, of $Pr\{X_{t+1}|X_t\}$ and $Pr\{s_t|X_t\}$, where t is the current time and X_t and s_t are the state and symbol (respectively) at time t.

Our interest is in generalizing the HMM approach so that we can investigate the usefulness of this model-class in discovering *de novo* underlying patterns in a long sequence of uncharacterized DNA. Our approach has two major facets:

(i) we generalize the HMM to a *non-deterministic probabilistic finite automaton* (NPFA) in which transitions among the hidden states are defined by the probabilities: $Pr\{X_{t+1} = i, s_t|X_t\} \geq 0$. (Non-determinism allows for more than one possible transition out of state X_t upon reading the symbol s_t.)

(ii) we infer the topology as well as the parameters of the NPFA from the data.

We are testing our methods using a simple stochastic finite automaton to generate a data stream. We have preliminary results for exon and intron data. Our hope is that we will discover structure in sequences that leads to insights into their fundamental biological properties.

Work supported under contract W-7405-ENG-36.

Sampling Based Methods for the Estimation of DNA Sequence Accuracy

Gary Churchill and Betty Lazareva
Biometrics Unit, Cornell University, Ithaca, NY 14853

We present a model for random errors that occur in DNA sequence data. The model is defined in terms of three parameters, one for each of the possible error types, substitution, insertion or deletion. A Gibbs sampling algorithm is described that can be used to simultaneously estimate the error rate parameters and to restore the DNA sequence. Parameter estimates are summarized as a posterior density. The restored DNA sequence can be summarized as a modal sequence or as a posterior credible region which takes the form of a cylinder set in sequence space. The methods are applied to a set of DNA sequence fragments from a human gene. Possible generalizations of the model and the algorithm are discussed in light of these results.

IDENTIFICATION, ORGANIZATION AND ANALYSIS OF MAMMALIAN REPETITIVE DNA

Jerzy Jurka, Genetic Information Research Institute, 1190 Eureka Ave., Los Altos, CA 94024

There are three major components of this project: organization of databases of mammalian repetitive sequences; development of specialized software for analysis of repetitive DNA and studies of new mammalian repeats. During the meeting we will report recent progress in all the three areas.

We will demonstrate recent development of the mammalian portion of the database of repetitive elements (repbase)[1]. The database is available electronically via internet (ftp ncbi.nlm.nih.gov; login: anonymous, password: your email address). Recent influx of sequence data to GenBank created unprecedented need for annotation of known repetitive elements. We will demonstrate our software for identification and elimination of repetitive DNA (CENSOR)[2], and a new software for Alu subfamily classification, based on the most recent progress in the field[3,4].

During the last few years over 40 new repetitive families have been reported in the human genome alone. Some of the new repeats turned out to be fragments of ancient L1 subfamilies[5]. This significantly improved our understanding of relationships between different repeats in the human genome. Other GenBank sequence studies revealed very abundant but ancient repeats called MIRs[6]. During the meeting we will report recently discovered families of MEdium Reiteration frequency repeats[7].

Supported by the U.S. Department of Energy, Office of Health and Environmental Research, grants DE-FG03-911ER61152 and DE-FG03-95ER62139.

[1]J. Jurka, Databases of repetitive elements (repbase). NCBI Database Repository: 1993, 1994 and 1995.
[2]J. Jurka, P. Klonowski, V. Dagman and P. Pelton, CENSOR - a program for identification and elimination of repetitive elements from DNA sequences, *Computers and Chemistry - special issue*, in press
[3]J. Jurka, *The Impact of Short Interspersed Elements (SINEs) on the Host Genome*, ed. Richard J. Maraia, R.G. Landes Company, pp. 25-41 (1995).
[4]V. Kapitonov and J. Jurka, The Age of Alu Subfamilies, *J. Mol. Evol.*, in press.
[5]A.F.A. Smit, G. Tóth, A.D. Riggs and J. Jurka, *J. Mol. Biol.* **246**, 401-417(1995).
[6]J. Jurka, E. Zietkiewicz and D. Labuda, *Nucl. Acids Res.* **23**, 170-175 (1995).
[7]J. Jurka, V. Kapitonov, A.F.A Smit, P. Klonowski, Identification of new medium reiteration frequency repeats in the human, rodent and rabbit genomes, in preparation.

NOVEL NEURAL NETWORKS FOR FUNCTIONAL SITE PREDICTION ON LARGE SCALE SEQUENCES*

Martin G. Reese, Nomi L. Harris and Frank H. Eeckman, Human Genome Informatics Group, Lawrence Berkeley National Laboratory, Berkeley CA 94720, mgreese@lbl.gov.

Recently several groups, including the LBNL Human Genome Center, scaled up the production efforts in human genome sequencing. One of the computational challenges in these projects is the automated detection of significant features in genomic DNA. We are especially interested in the recognition of functional sites like promoter elements, coding regions, and splice site junctions. We believe that better detection of these features will lead to vastly improved gene finding. However, promoter elements and splice junctions have a complex structure, consisting of many individual elements, such as the TATA-box, transcription start signal for promoters and splice site consensus sequences. Furthermore, the relative positions of the individual elements are variable, and some elements may be absent altogether. Previous efforts in this area have been plagued by high rates of false positives. Given the nature of genomic sequencing in humans, where large introns are known to exist, we recognize the need for a very specific algorithm, with a small number of false positives.

To predict functional elements and regions in DNA sequences we use neural networks which tend to predict to a very high degree of selectivity.

For the promoter prediction problem we use a special architecture of a neural network, a time-delay neural network, to combine the predictions that were made for each of the individual promoter elements. TDNNs are appropriate for recognizing promoter elements because they are able to combine multiple features, even those that appear at different relative positions in different sequences. Another advantage is the high selectivity of the TDNN, which is extremely important for promoter prediction systems, in order to avoid generating too many false positives.

Our TDNN predicts most of the annotated promoters in a set of human genes from Genbank (version 86.0). As an example, the TDNN finds the annotated promoter from a 13,865 basepair test gene, HUMTFPB, with a false positive score of 0.05% (6 false positive predictions out of 13,865).

We have applied this network and the splice site prediction networks to our most recently produced human DNA sequences and will present data at the conference.

* This work was supported by the U.S. Department of Energy under Contract Number DE-AC03-76SF00098

Languages, Automata, Interfaces, and Macromolecules*

David B. Searls, SmithKline Beecham Pharmaceuticals, 709 Swedeland Road, PO Box 1539, King of Prussia, PA 19406

Viewed as strings of symbols, biological macromolecules can be modelled as elements of formal languages. Generative grammars have been useful in molecular biology for purposes of syntactic pattern recognition, for example in the author's work on the GenLang pattern matching system, which is able to describe and detect patterns that are provably beyond the capability of a regular expression specification. More recently, grammars have been used to capture intramolecular interactions or long-distance dependencies between residues, such as those arising in folded structures. In the work of Haussler and colleagues, for example, stochastic context-free grammars have been used as a framework for "learning" folded RNA structures such as tRNAs, capturing both primary sequence information and secondary structural covariation. Such advances make the study of the formal status of the language of biological macromolecules highly relevant, and in particular the finding that DNA is beyond context-free has already created challenges in algorithm design.

Moreover, to date such methods have not been able to capture relationships between strings in a collection, such as those that arise via intermolecular interactions, or evolutionary relationships implicit in alignments. Recently we have attempted to remedy this by showing (1) how formal grammars can be extended to describe interacting collections of molecules, such as hybridization products and, potentially, multimeric or physiological protein interactions, and (2) how simple automata can be used to model evolutionary relationships in such a way that complex model-based alignment algorithms can be automatically generated by means of visual programming. These results allow for a useful generalization of the language-theoretic methods now applied to single molecules.

In addition, we describe a new software package for genome application development, called bioTk. This is a domain-specific widget set, implemented in Tcl/Tk, for the rapid prototyping of sophisticated graphical user interfaces involving objects such as chromosomes, maps, and sequences.

*Supported by a grant from the Department of Energy, 92ER61371.

D.B. Searls, "String Variable Grammar: A Logic Grammar Formalism for DNA Sequences", *Journal of Logic Programming* **24**(1,2):73-102 (1995).

D.B. Searls, "Formal Grammars for Intermolecular Structure", *First International Symposium on Intelligence in Neural and Biological Systems,* 30-37 (1995).

D.B. Searls and K.P. Murphy, "Automata-Theoretic Models of Mutationa and Alignment", *Third International Conference on Intelligent Systems for Molecular Biology,* 341-349 (1995).

D.B. Searls, "bioTk: Componentry for Genome Informatics Graphical User Interfaces", *Gene* **163**(2):GC1-16 (1995).

Algorithms in Support of the Human Genome Project[1]

Dan Gusfield, Jim Knight, Kevin Murphy, Paul Stelling Lushen Wang,

Department of Computer Science, University of California, Davis, CA 95616. gusfield@cs.ucdavis.edu
And Archie Cobbs, Paul Horton, Gene Lawler

Department of Computer Science, University of California, Berkeley, CA

Our research covers a wide variety of algorithmic and data structure issues involved in obtaining and analyzing sequence data, in searching databases, in reconstructing sequences from hybridization data, in reconstructing evolutionary history from sequence data or from genome rearrangements, in studying repeated structures in biological sequences. The work is both theoretical and applied, and has produced more than twenty papers and five computer programs in the last two years. Below is a selected listing of some of the recent efforts supported by the grant.

Sequence analysis and database searching:

Fast identification of approximately matching substrings - Cobbs Published in the Proceedings of the 6'th Annual symposium on combinatorial pattern matching, July 1995.

Improved approximate matching over suffix trees - Cobbs Published in the Proceedings fo the 5'th Annual symposium on combinatorial pattern matching, June 1994.

Uniform preprocessing for linear time string matching - Gusfield Dimacs Technical report, December 1995.

Approximate algorithms for multiple sequence alignment - Bafna, Lawler, Pevzner Published in the Proceedings of the 5'th Annual symposium on combinatorial pattern matching, June 1994.

Computational experience with a branch-and-bound algorithm for maximum-trace multiple sequence alignment - Kececioglu Published in the Proceedings of the 5'th Annual symposium on combinatorial pattern matching, June 1994.

Automata-Theoretic Models of Mutation and Alignment - David Searls and Kevin Murphy Published in the Proceedings of the 3'd International conference on Intelligent Systems in Molecular Biology. July, 1995.

Efficient Parametric and Inverse Parametric Sequence Alignment with XPARAL – Gusfield and Stelling. To appear in Methods of Enzymology issue on Computer Methods for Macromolecular Sequence Analysis edited by R. Doolittle.

A Branch and Bound Algorithm for Local Multiple Alignment - Horton. To appear in the Proceedings of the Pacific Symposium on Biocomputing January 1996.

Prioritized Suffix Tree Based Approximate Search - Cobbs Manuscript June 1995.

Constructing Additive Trees when the Error is Small - Wang Manuscript, August 1995.

An Efficiently Solved Travelling Salesman Problem arising from Sequencing by Hybridization. Gusfield, Stelling, Wang. Manuscript, October 1995.

Optimal Alignments in Linear Space Using Automaton-derived Cost Functions - Murphy. Manuscript July 1995.

Passively Learning Finite Automata: A Survey - Murphy. Manuscript, October 1995.

Genome Rearangements:

Efficient Bound for oriented chromosome-inversion distance - Kececioglu, Sankoff In Proceedings of the 5th *Symposium on Combinatorial Pattern Matching*, June 1994.

Of Mice and Men: Algorithms for evolutionary distances between genomes with translocation and inversion - Kececioglu and Ravi Published in the Proceedings of the 6th ACM-SIAM *Symposium on Discrete Algorithms*, January 1995.

Sequence Reconstruction

Approximate algorithms for multiple alignment to a phylogenetic tree - Jiang, Lawler, Wang. To appear in SIAM Journal on Computing.

Improved Algorithms for Tree Alignment - Wang and Gusfield Manuscript, October 1995.

Approximation algorithms for multiple sequence alignment under a fixed evolutionary tree – Ravi and Kececioglu. Proceedings of the 6'th Annual symposium on combinatorial pattern matching, July, 1995.

[1]Supported by DOE Grant DE-FG03-90ER60999

A NEW TECHNIQUE FOR DETECTING CODING DNA SEQUENCES†

David C. Torney, Clive C. Whittaker, and Guochun Xie, Center for Human Genome Studies, MSK710, Los Alamos National Laboratory, Los Alamos, NM 87545

Improved detection of human coding sequences is highly desirable for many reasons, including SAmple SEquencing (see Chi *et al.* and Ricke *et al.* this meeting). We developed a novel approach which, in principle, makes full use of all differences between coding and noncoding sequence data for classification. Likelihoods of sequence data lie at the root of this approach. We first converted the DNA sequences to binary sequences, encoding as follows: A=00, C=01, G=10, T=11. Let the *parity* of a binary sequence be the number of ones modulo 2. For clarity, begin with two datasets of example sequences, coding and noncoding, with n (binary) letters in each sequence[1]. Count the number of times the parity is even for subsequences– not necessarily consecutive subsequences. Although there are 2^n subsequences, it is natural to focus on those subsequences with the smallest number of letters. In fact, the distribution of the differences of the average parity of a subsequence between the two datasets narrowed as the number of letters increased. Our motivation was completeness, and it should be noted that our approach relies upon none of the mainstays of other techniques, such as subsequence frequencies or periodicities[1,2].

To establish feasibility, we considered only those subsequences of up to six (binary) letters, and we required the first and last letters to be within 60 letters of one another. Neither the phase nor the strand of the coding sequences were known in the "training" dataset[1]; to partially mitigate the latter we appended the reverse-complementary sequences. Lacking the phase, we averaged the parities of subsequences which were translates (modulo 2) of one another. To classify "test" sequences, we retained only a small number of subsequences: essentially those with the largest magnitude of the difference in the average parities for the "training" coding and noncoding data. Retained subsequences frequently had some pairs of letters corresponding to individual bases, but a three-base periodicity was uncommon. The subsequence with two adjacent letters, corresponding to an individual base, had the largest magnitude of the difference in average parities between coding and noncoding sequences, reflecting the larger C + G content of coding sequences.

Finally, for "test" sequences, we added the retained subsequences' parities, each multiplied by the difference of the average parities in the training datasets. A threshold was used to classify: for sums above the threshold the classification was noncoding. The threshold was chosen to make the false-prediction rates in the two test datasets equal. After analyzing approximately 72,000 54-base training and test sequences, the false-prediction rates we obtained were 27.5%, whereas 29.5% was the smallest previously found for the same dataset using an individual feature: hexamer frequencies[1]. Because of the restrictions in our preliminary feasibility studies, substantial improvements are likely. Our approach might also contribute to modular prediction software, such as GRAIL.

†This work was funded by the U.S. D.O.E. under contract W-7405-ENG-36.

1. J. W. Fickett and C.-S. Tung, *Nucleic Acids Res.*,**20**, 6441-6450, (1992).
2. A. Thomas and M. H. Skolnick, *IMA Journal of Mathematics Applied in Medicine and Biology*, **11**, 149-160 (1994).

Gene Recognition, Modeling, and Homology Search in GRAIL and genQuest

Ying Xu[1], Manesh Shah[1], J. Ralph Einstein[1], Sherri Matis[1], Xiaojun Guan[1], Sergey Petrov[1], Loren Hauser[2], Richard J. Mural[2], and Edward C. Uberbacher[1]
[1]Computer Science and Mathematics, and [2]Biology Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6364. e-mail:GRAILMAIL@ornl.gov

GRAIL is a modular expert system for the analysis and characterization of DNA sequences which facilitates the recognition of gene features and gene modeling. A new version of the system has been created with greater sensitivity for exon prediction (especially in AT rich regions), more accurate splice site prediction, and robust indel error detection capability. GRAIL 1.3 is available to the user in a Motif graphical client-server system (XGRAIL), through WWW-Netscape, by email server, or callable from other analysis programs using Unix sockets.

In addition to the positions of protein coding regions and gene models, the user can view the positions of a number of other features including poly-A addition sites, potential Pol II promoters, CpG islands and both complex and simple repetitive DNA elements using algorithms developed at ORNL. XGRAIL also has a direct link to the genQuest server, allowing characterization of newly obtained sequences by homology-based methods using a number of protein, DNA, and motif databases and comparison methods such as FastA, BLAST, parallel Smith-Waterman, and special algorithms which consider potential frameshifts during sequence comparison.

Following an analysis session, the user can use an annotation tool which is part of the XGRAIL 1.3 system to generate a "feature table" report describing the current sequence and its properties. Links to the GSDB sequence database have been established to record computer-based analysis of sequences during submission to the database or as third party annotation.

Gene Modeling and Client-Server GRAIL: In addition to the current coding region recognition capabilities based on a multiple sensor-neural network and rule base, modules for the recognition of features such as splice junctions, transcription and translation start and stop, and other control regions have been constructed and incorporated into an expert system (GAP III) for reliable computer-based modeling of genes. Heuristic methods and dynamic programming are used to construct first pass gene models which include the potential for modification of initially predicted exons. These actions result in a net improvement in gene characterization, particularly in the recognition of very short coding regions. Translation of gene models and database searches are also supported through access to the *genQuest* server (described below).

Model Organism Systems: A number of model organism systems have been designed and implemented and can be accessed within the XGRAIL 1.3 client including *Escherichia coli, Drosophila melanogaster* and *Arabidopsis thaliana.* The performance of these systems is basically equivalent to the Human GRAIL 1.3 system. Additional model organism systems, including several important microorganisms, are in progress.

Error Detection in Coding Sequences: Single-pass DNA sequencing is becoming a widely used technique for gene identification from both cDNA and genomic DNA sequences. An appreciably higher rate of base insertion and deletion errors (indels) in this type of sequence can cause serious problems in the recognition of coding regions, homology search, and other aspects of sequence interpretation. We have developed two error detection and "correction" strategies and systems which make low-redundancy sequence data more informative for gene identification and characterization purposes. The first algorithm detects sequencing errors by finding changes in the statistically preferred reading frame within a possible coding region and then rectifies the frame at the transition point to make the potential exon candidate frame-consistent. We have incorporated this system in GRAIL 1.3 to provide analysis which is very error

tolerant. Currently the system can detect about 70% of the indels with an indel rate of 1%, and GRAIL identifies 89% of the coding nucleotides compared to 69% for the system without error correction. The algorithm uses dynamic programming and runs in time and space linear to the size of the input sequence.

In the second method, a Smith-Waterman type comparison is facilitated in which the frame of DNA translation to protein sequence can change within the sequence. The transition points in the translation frame are determined during the comparison process and a best match to potential protein homologs is obtained with sections of translations from more than one frame. The algorithm can detect homologies with a sensitivity equivalent to Smith-Waterman in the presence of 5% indel errors.

Detection of Regulatory Regions: An initial Polymerase II promoter detection system has been implemented which combines individual detectors for TATA, CAAT, GC, cap, and translation start elements and distance information using a neural network. This system finds about 67% of TATA containing promoters with a false positive rate of one per 35 kilobases. Additionally a systems to detect potential polyA addition sites and CpG islands has been incorporated into GRAIL

The GenQuest Sequence Comparison Server: The genQuest server is an integrated sequence comparison server which can be accessed via e-mail, using Unix sockets from other applications, Netscape, and through a Motif graphical client-server system. The basic purpose of the server system is to facilitate rapid and sensitive comparison of DNA and protein sequences to existing DNA, protein, and motif databases. Databases accessed by this system include the daily updated GSDB DNA sequence database, SwissProt, the dbEST expressed sequence tag database, protein motif libraries and motif analysis systems (Prosite, BLOCKS), a repetitive DNA library (from J. Jurka), Genpept, and sequences in the PDB protein structural database. These options can also be accessed from the **XGRAIL** graphical client tool.

The genQuest server supports a variety of sequence query types. For searching protein databases, queries may be sent as amino acid or DNA sequence. DNA sequence can be translated in a user specified frame or in all 6 frames. DNA-DNA searches are also supported. User selectable methods for comparison include the Smith-Waterman dynamic programming algorithm, FastA, versions of BLAST, and the IBM dFLASH protein sequence comparison algorithm. A variety of options for search can be specified including gap penalties and option switches for Smith-Waterman, FastA, and BLAST, the number of alignments and scores to be reported, desired target databases for query, choice of PAM and Blosum matrices, and an option for masking out repetitive elements. Multiple target databases can be accessed within a single query.

Additional Interfaces and Access: Batch GRAIL 1.3 is a new "batch" GRAIL client allows users to analyze groups of short (300-400 bp) sequences for coding character and automates a wide choice of database searches for homology and motifs. A **Command Line Sockets Client** has been constructed which allows remote programs to call all the basic analysis services provided by the GRAIL-genQuest system without the need to use the XGRAIL interface. This allows convenient integration of selected GRAIL analyses into automated analysis pipelines being constructed at some genome centers. An **XGRAIL Motif Graphical Client** for the GRAIL release 1.3 has been constructed using Motif with versions for a wide variety of UNIX platforms including Sun, Dec, and SGI. The e-mail version of GRAIL can be accessed at *grail@ornl.gov* and the e-mail version of genQuest can be accessed at *Q@ornl.gov*. Instructions can be obtained by sending the word "help" to either address. The Motif or Sun versions of XGRAIL, batch GRAIL, and XgenQuest client software are available by anonymous ftp from arthur.epm.ornl.gov (128.219.9.76). Both GRAIL and genQuest are accessible over the World Wide Web (URL http://avalon.epm.ornl.gov/). Communications with the GRAIL staff should be addressed to *GRAILMAIL@ornl.gov*. (Supported by the Office of Health and Environmental Research, United States Department of Energy, under contract DE-AC05-84OR21400 with Lockheed Martin Energy Systems,Inc.)

Prediction of Coding Regions and Promoters in Genomic DNA*

Gary D. Stormo

Department of Molecular, Cellular and Developmental Biology
University of Colorado, Boulder, CO 80309

We have developed an approach for predicting coding regions in genomic DNA that utilizes multiple types of evidence, combines those into a single scoring function and then returns both optimal and ranked suboptimal solutions using that scoring function.[1] By including separate scoring functions for loci with different G+C content, the method improves prediction overall, and especially for the usually difficult low G+C genes. The use of similarity matches, in the form of BLAST "hits" further increases the reliability of the predictions considerably. Alternative splicing pathways often show up in the suboptimal plots. The approach is shown to be robust to substitution errors in the sequence, but highly susceptible to frame-shift errors. The approach can easily be extended to other problems where a sequence is to be partitioned into domains belonging to a set of possible functional classes. It can also be modified such that the probability of the correct parsing is maximized over a training set of examples.[2] This modification allows for the use of a stochastic grammar to describe the class of possible parses, and for the weighting of various types of evidence to be adjusted to obtain the highest reliability.

We are now exploring methods for reliably predicting other classes of sequence regions, especially promoters. These include approaches based on minimal length encoding algorithms and on Markov chains for the various classes. Some new advances have been made and will be described.

* Supported by a grant from the U.S. Department of Energy under contract ER61606.

[1] E.E. Snyder and G.D. Stormo, *J. Mol. Biol.* **248**, 1-18 (1995).

[2] G.D. Stormo and D. Haussler, *Proceedings of the Second International Conference on Intelligent Systems in Molecular Biology* , pp. 369-375 (1994).

Spliced Alignment: a New Approach to Gene Recognition

Mikhail S. Gelfand[1], Andrey A. Mironov[2], Pavel A. Pevzner[3]

[1] Institute of Protein Research, Russian Academy of Sciences,
Puschino, Moscow region, 142292, Russia.
misha@imb.imb.free.net

[2] Laboratory of Mathematical Methods, National Center for Biotechnology NIIGENETIKA,
Moscow, 113545, Russia.
mir@vnigen.msk.su

[3] Departments of Mathematics and Computer Science, University of Southern California,
Los Angeles, CA 90089-1113
pevzner@cse.psu.edu

Gene recognition is one of the most important problems in computational molecular biology. Previous attempts to solve this problem were based on statistics and artificial intelligence and, surprisingly enough, applications of theoretical computer science methods for gene recognition were almost unexplored. Recent advances in large-scale cDNA sequencing open a way towards a new combinatorial approach to gene recognition. This paper describes a *spliced alignment* algorithm and a software tool which explores all possible exon assemblies in polynomial time and finds the multi-exon structure with the best fit to a related protein. Unlike other existing methods, the algorithm successfully performs exons assemblies even in the case of short exons or exons with unusual codon usage; we also report correct assemblies for genes with more than 10 exons provided a homologous protein is already known. On a test sample of human genes with known mammalian relatives the average overlap between the predicted and the actual genes was 98%, which is remarkably well as compared to other existing methods. At that, the algorithm absolutely correctly reconstructed 87% of genes. The rare discrepancies between the predicted and real exon-intron structues were restricted either to extremely short initial or terminal exons (less than 5 amino acids) or proved to be results of alternative splicing and errors in database feature tables. Moreover, the algorithm performs reasonably well with non-vertebrate and even prokaryote targets. The dependence of the performance on the evolutionary distance between the analyzed gene and the target protein was estimated using simulated data. The main result of the paper is that a relatively simple algorithm based on combinatorial common sense and some biological intuition can outperform many approaches developed for gene recognition in the last fifteen years.

This work is supported by DOE grant DE-FG02-94ER61919.

Fast Protein Folding in the Hydrophobic-hydrophilic Model Within Three-eighths of Optimal[1]

William E. Hart and Sorin Istrail
Sandia National Laboratories
MS 1110, Albuquerque, NM 87185-1110

We present performance-guaranteed approximation algorithms for the protein folding problem in the hydrophobic-hydrophilic model, Dill (1985)[1]. To our knowledge, our algorithms are the first approximation algorithms in the literature with guaranteed performance for this model, Dill (1994)[2]. The hydrophobic-hydrophilic model abstracts the dominant force of protein folding: the hydrophobic interaction. The protein is modeled as a chain of amino acids of length n that are of two types: H (hydrophobic, i.e., nonpolar) and P (hydrophilic, i.e., polar). Although this model is a simplification of more complex protein folding models, the protein folding structure prediction problem is notoriously difficult for this model. Our algorithms have linear ($3n$) or quadratic time and achieve a three-dimensional protein conformation that has a guaranteed free energy no worse than $\frac{3}{8}$ of optimal. This result answers the open problem of Ngo, Marks and Karplus (1994)[3] about the possible existence of an efficient approximation algorithm with guaranteed performance for protein structure prediction in any well-studied model of protein folding. By achieving speed and near-optimality simultaneously, our algorithms rigorously capture salient features of the recently proposed framework of protein folding by Sali, Shakhnovich and Karplus (1994)[4]. Equally important, the final conformations of our algorithms have significant secondary structure (anti-parallel sheets, beta sheets, compact hydrophobic core). Furthermore, hypothetical folding pathways can be described for our algorithms that fit within the framework of diffusion-collision protein folding proposed by Karplus and Weaver (1979)[5]. Computational limitations of algorithms that compute the optimal conformation have restricted their applicability to short sequences (length less than or equal to 90). Because our algorithms trade computational accuracy for speed, they can construct near-optimal conformations in linear time for sequences of any size.

References

[1] K. A. Dill, *Biochemistry*, 24:1501, 1985

[2] K. A. Dill, Personal Communication, 1994

[3] J. T. Ngo, J. Marks, M. Karplus, *Computational complexity, protein structure prediction, and the Levinthal paradox* in *The Protein Folding Problem, ch. 14, pp. 435-508*, Birkhauser, 1994

[4] A. Sali, E. Shakhnovich, M. Karplus, *Nature*, 369:248-251, 1994

[5] M. Karplus, D. L. Weaver, *Biopolymers* 18:1421, 1979

[1]This work was supported by the U.S. Department of Energy and was performed at Sandia National Laboratories for the U.S. Department of Energy under contract DE-AC04-94AL85000.

Smith-Waterman Transformable Array Search Engine Yields Supercomputer Performance on PCs and Unix Workstations

James W. Lindelien, Robert Farrington, Betty Tjandra, Time Logic, Inc., 11992 Challenger Ct., Moorpark, CA 93021; and 567 Knotty Pine Dr., Incline Village, NV 89451 e-mail to **jiml@sierra.net**

With government funding incentives to reduce the cost per sequenced base an order of magnitude by the year 2000, the informatics bottleneck must be addressed economically or further scale-up of massive sequencing projects will be significantly hindered. We have developed a transformable array accelerator based on Field-Programmable-Logic-Array (FPGA) technology for high-speed searches of massive nucleic and protein databases that is one to two orders of magnitude less expensive than present systems, yet yields the world's fastest Smith-Waterman implementation. We hope to promote discovery by allowing more analysis per research dollar, by more biologists. The ability to retrieve search results in seconds permits a more creative "what-if" style of interactive database exploration, versus conventional but tedious "batch" style searches. System performance is scaleable over 100x.

Multiple search processors are configured within an on-card FPGA array. Less computationally intensive algorithms run up to 10x faster than Smith-Waterman due to increased on-chip parallelism. Since the on-chip wiring of FPGAs can be set by software configuration in a fraction of a second, the array is readily transformed for a variety of search algorithms, or to first "pre-screen" large query sets with a higher speed (but less sensitive) algorithm to yield a higher daily throughput. The algorithms are:

ALGORITHM	PARAMETERS	PERFORMANCE/card
Nucleotide-to-Nucleotide comparisons, ungapped	IUB/GCG-16 Symbol Set; ktup=1. Dual scoring: k-tuple length, and max. % matches in user set "bin" size of 16 to 512 bases/bin (powers of 2).	**1.28 Billion nt-nt comparisons/sec** (ktup=1)
Amino Acid-to-Amino Acid comparisons, ungapped	Using 5 bit 32x32 programmable matrix; ktup=1. Dual scoring: k-tuple length, and max. % matches in user set "bin" size of 16 to 512 residues/bin (powers of 2).	**320 Million aa-aa comparisons/sec** (ktup=1)
Smith-Waterman	User programmable similarity matrix; affine gap scoring to 16 bit resolution.	**100 Million S-W matrix cells/sec**
Profile Search	User programmable similarity matrix; affine gap scoring to 16 bit resolution.	**100 Million matrix cells/sec**
Smith-Waterman Frame-Shift Tolerant	User programmable similarity matrix; and affine gap scoring. Searches 3 nt frames vs. aa target simultaneously with user set frame shift and gap penalties.	**100 Million S-W matrix cells/sec**
Profile Search Frame-Shift Tolerant	User programmable similarity matrix; and affine gap scoring. Searches 3 nt frames vs. aa target simultaneously with user set frame shift and gap penalties.	**100 Million matrix cells/sec**

For high-volume sequencing projects, the multiple query, single target design accepts in bulk the query sets produced by automated sequencers. Simultaneous processing of many short queries creates a larger "effective query size." This decouples the computation rate of the FPGA array from the disk read bandwidth, permitting very high performance without a disk bottleneck, nor the usual requirement to RAM-cache the target database. Typical Pentium-based computers with low-cost disk systems (EIDE, SCSI-2) support Smith-Waterman searches on up to 15 accelerator cards per PC, at an aggregate rate of 1.5 billion SW cells/second. Such a PC is about 110% the speed of the MasPar MP-2 16384 processor supercomputer, but costs only 4% as much, and requires no annual support contract. Still higher performance is achieved by TCP/IP (NFS) network clustering. We offer individual cards and development assistance to academic researchers having other applications for the technology.

This project was privately funded by Time Logic, Inc.

This page intentionally left blank.

Ethical, Legal, and Social Issues

This page intentionally left blank.

THE HUMAN GENOME: SCIENCE AND THE SOCIAL CONSEQUENCES; INTERACTIVE EXHIBITS AND PROGRAMS ON GENETICS AND THE HUMAN GENOME

Charles C. Carlson, Director, Life Sciences
The Exploratorium; San Francisco, CA 94123
415/561-0319, FAX 415/561-0307; Internet: *charliec@exploratorium.edu*

From April through September, 1995, the Exploratorium mounted a special exhibition called *Diving into the Gene Pool* consisting of 26 interactive exhibits developed over the course of three years. The exhibits introduce the science of genetics and increase public awareness of the Human Genome Project and its implications for society. Founded in the success of exhibits developed for the 1992 genetics and biotechnology symposium "Winding Your Way Through DNA" (co-hosted with the University of California, San Francisco), the 1995 exhibition aimed to create an engaging and accessible presentation of specific information about genetic science and our understanding of the structure and function of the human genome, genetic technology, and ethical issues surrounding current genetic science.

In addition to creating a unique collection of exhibits, the project developed a range of supplemental public programming to provide public forum for discussion and interaction about genetics and bioethics. A lecture series entitled "Bioethics and the Human Genome Project," featured such key thinkers as Mary Claire King, Leroy Hood, David Martin, Troy Duster, Michael Yesley, William Atchley, and Joan Hamilton (among others). A weekend event program focused on biodiversity in animal and plant life with events such as "Seedy Science," "Blooming Genes," and "Dog Diversity." A Biotech Weekend offered access to new technologies through demonstrations by local biotech firms and genetic counselors. And a specially-commissioned theatre piece, "Dog Tails," provided a instructive and comic look for kids into the foundations of genetics and issues of diversity.

In the 5-month exhibition period, approximately 300,000 visitors had the opportunity to visit the exhibition, and well over 5,000 participated in the special programming. Following the exhibition's close, the new exhibits will become a permanent part of the Exploratorium's collection of over 650 interactive exhibits.

Additional funding for 1995-96 will suport formal outside evaluation of the effectiveness of the exhibits, and support exhibit remediation based on the evaluation findings. This activity will both strengthen the Exploratorium's permanent collection of genetics exhibits and help to develop a feasibility study for a travelling version of the genetics exhibition for other museums around the country and the world.

Supported by a grant from the Program Director, U. S. Department of Energy-ELSI under grant number DE-FG03-93ER61583.

COMMUNICATING SCIENCE IN PLAIN LANGUAGE: THE SCIENCE + LITERACY FOR HEALTH: HUMAN GENOME PROJECT*

Maria Sosa, Judy Kass, and Tracy Gath, American Association for the Advancement of Science. 1333 H Street, NW, Washington, DC 20005.

Recent literacy surveys have found that a large number of adults lack the skills to bring meaning to much of what is written about science. This, in effect, denies them access to vital information about their health and well-being. To address this need, the American Association for the Advancement of Science (AAAS) is developing a 2-year project to provide low-literate adults with the background knowledge necessary to address the social, ethical, and legal implications of the Human Genome Project.

With its **Science + Literacy for Health: Human Genome Project**, AAAS is using its existing network of adult education providers and volunteer science and health professionals to pursue the following overall objectives: (1) to develop new materials for adult literacy classes, including a high-interest reading book and accompanying curriculum, an implementation framework, a short video providing background information on genetics, a database of resources, and fact sheets that will assist other organizations and researchers in preparing easy-to-read materials about the human genome project, and (2) to develop and conduct a campaign to disseminate project materials to libraries and community organizations carrying out literacy programs throughout the United States.

Because not every low-literate adult is enrolled in a literacy class, our model for helping scientists communicate in simple language will have impact beyond classrooms and learning centers. In preliminary contacts, community groups providing health services have indicated that the proposed materials are not only desirable but needed; indeed such groups often receive requests for information on heredity and genetics. The module developed by AAAS should enable other medical and scientific organizations to communicate more effectively with economically disadvantaged populations, which often include a large number of low-literate individuals.

* Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy under contract DE-FG02-95ER61988.

THE DNA FILES: A NATIONALLY SYNDICATED SERIES OF RADIO PROGRAMS ON THE SOCIAL IMPLICATIONS OF HUMAN GENOME RESEARCH AND ITS APPLICATIONS*

Bari Scott, Matt Binder, and Jude Thilman, Genome Radio Project, produced at KPFA-FM, Berkeley, CA 94704.

The DNA Files is a series of nationally distributed public radio programs furthering public education on developments in genetic science. Program content is guided by a distinguished body of advisors and will include the voices of prominent genetic researchers, people affected by advances in the clinical application of genetic medicine, members of the biotech industry, and others from related fields. They will provide real-life examples of the complex social and ethical issues associated with new discoveries in genetics. In addition to the general public radio audience, the series will target educators, scientists, and involved professionals. Ancillary educational materials will be distributed in paper and digital form through over two dozen collaborative organizations and fulfillment of listener requests.

"DNA and Behavior: Is Our Fate Written in Our Genes?" is the pilot documentary for the series, scheduled for release in early 1996. The show will help the lay person understand and evaluate recent research in the area of behavioral genetics. Recently, we've seen news media reports on newly discovered genetic factors being related to behaviors such as alcoholism, mental illness, sexual orientation and aggression. This program will look at several examples of these "genetic factors" and evaluate the strengths and weaknesses of various methodologies involved in the research; and introduce such controversial issues as the re-emergence of a eugenics movement based on theoretical suppositions drawn from recent work in behavioral genetics.

With information linking major diseases such as breast cancer, colon cancer, and arteriosclerosis to genetic factors, new dangers in public perception emerge. Many people who hear about them mistakenly conclude that these diseases can now be easily diagnosed and even cured. On the other end of the public perception spectrum, unfounded fears of extreme, and highly unlikely, consequences also appear. Will society now genetically engineer whole generations of people with "designer genes" offering more "desirable physical qualities"? *The DNA Files* will ground public understanding of these issues in reality. "DNA and the Law" reviews the scientific basis for genetic fingerprinting and looks at cases of alleged genetic discrimination by insurance companies, employers and others. This program also looks at disputes over paternity, intellectual property rights, the commercialization of genetic information, informed consent and privacy issues. Other shows include "The Search for a Breast Cancer Gene," "Prenatal Genetic Testing and Treatment," "Evolution and Genetic Diversity," "Sickle-Cell Disease and Thalassemia: Hope for a Cure," and "Theology, Mythology and Human Genetic Research."

*Supported by ELSI grant DE-FG03-95ER62003 from the Office of Health and Environmental Research of the U.S. Department of Energy.

THE GENE LETTER:
A Newsletter on Ethical, Legal, and Social Issues in Genetics for Interested Professionals and Consumers

Philip Reilly, Dorothy Wertz and Robin Blatt

As the Human Genome Project continues to generate an ever more dense map, the pace of localization and cloning of genes that contribute to human disease grows. During the mid-nineties genomics has emerged as a major component of biotechnology and the stream of online DNA based diagnostic tests has become a small river. There can no longer be any doubt that we are entering an era in which we will be able to generate significant amounts of genetic data about individuals that may be relevant to their health. This explosive growth in our ability to acquire genetic information raises many challenging questions concerning how it could and should be used. During the first five years of the HGP the ELSI working group fostered much discussion and awareness among professionals about the potential for misuse of genetic data by insurers and employers. In the second five years of the HGP there is an obvious need to extend the boundaries of discourse to educate as many persons as possible concerning potential problems raised by advances in our understanding of the genome and to involve them in generating responsible social guidelines.

THE GENE LETTER offers one means to accomplish this task. This electronic newsletter will be published monthly on the Internet beginning in late Spring of 1996. Features will be largely informational and will include reports of new scientific/medical developments and attendant ELSI issues, new court decisions, legislation, and regulations, and comments about the manner in which such topics are reported in the traditional media. Some reports will be accompanied by analysis and/or editorial comments. An editorial board will review each issue prior to publication to check for accuracy,fairness, balance, adequate attention to the concerns of persons with disabilities and for cultural sensitivity. Readers will be invited to communicate with the authors either via the THE GENE LETTER Internet chatroom or by hard copy. Informal polls of readers concerning their views of various ethical issues will be conducted and reported. Important breaking news will be reported in a flash format to be followed up with longer commentary in the next issue.

The authors, a clinical geneticist/attorney, a social scientist, and a nurse, all have extensive experience in educating health care providers and the general public about advances in genetics and concomitant social issues.

THE HUMAN GENOME PROJECT AND MENTAL RETARDATION AN EDUCATIONAL PROGRAM

Sharon Davis, Ph.D., Director, Department of Research and Program Services, The Arc of the United States, 500 East Border St., Suite 300, Arlington, TX 76010

The Arc of the United States, a national organization on mental retardation, with 140,000 members and more than 1000 affiliated chapters proposes to educate its general membership and volunteer leaders about the Human Genome Project as it relates to mental retardation. A large number of identified causes of mental retardation are genetic, and many family members of The Arc deal with issues related to a genetic condition on a daily basis. We believe it is critical for our members and leaders to be educated about the scientific and ethical, legal and social aspects of the HGP, so that the association can evaluate and discuss the issues and develop positions based on adequate knowledge.

The major objectives of the proposed three-year project are to develop and disseminate educational materials for members/leaders of The Arc to inform them about the Human Genome Project and mental retardation and to conduct training on the scientific and ethical, legal and social aspects of the Human Genome Project and mental retardation using The Arc's existing training vehicles.

The Arc will develop and disseminate educational materials oriented toward families and conduct training at its national and state conventions, local chapter meetings and at board of director's meetings. The American Association of University Affiliated Programs for Persons with Developmental Disabilities (AAUAP) will assist with the project by providing needed expertise. The AAUAP membership includes university faculty who are experts on the genetic causes of mental retardation and on related ethical, legal and social issues. An advisory panel of university scientists and leaders of The Arc will guide the project.

HUMAN GENOME TEACHER NETWORKING PROJECT*

Debra L. Collins, and R. Neil Schimke, Genetics Education Center, Division of Endocrinology and Genetics, University of Kansas Medical Center, Kansas City, KS 66160-7318

This project links over 150 middle and secondary teachers from throughout the United States with genetic and public policy professionals, as well as families who are knowledgeable about the ethical, legal, and social implications (ELSI) of the Human Genome Project. Teachers network with peers and professionals, and acquire new sources of information during four phases: 1) the first one-week summer workshop to update teachers on human genetics concepts and new sources for classroom curricula including online resources; 2) classroom use of new materials and information; 3) the second one-week summer workshop where teachers return to exchange successful teaching ideas and plan peer teaching sessions and mentor networking; 4) dissemination of genetic information through in-services and workshops for colleagues; and collaboration with genetic professional participating in our *Mentor Network*.

The applications of Human Genome Project technology are emphasized. Individuals who have contact and experience with patients, including clinical geneticists, genetic counselors, attorneys, laboratories geneticists and families, take part in didactic sessions with teachers. Throughout the workshop, *family panels* provide an opportunity for participants to compare their textbook-based knowledge of genetic conditions with the personal experiences of families who discuss their condition, including: diagnosis, treatment, genetic risk, decisions, insurance, employment, family planning, and confidentiality.

Because of this project, teachers feel more prepared and confident teaching about human genetics, the Human Genome Project, and ELSI topics. The teachers are effective in disseminating knowledge of genetics to their students who show a significant increase in human genome knowledge compared to students whose teachers have not participated in this project.

Teacher dissemination activities extend the project beyond participation at summer workshops. To date, 55 workshop participants have completed all four project phases by organizing more than 200 local, regional, and national teacher education programs to disseminate knowledge and resources. More than 9,500 colleagues and the general public have participated in teacher workshops, and over 560,00 students have been reached through project participants and their peers.

The project participants organize interdisciplinary peer teaching sessions including bioethical decision making sessions combining debate and biology classes; sessions for social studies teachers; human genetics and multi-cultural collaborations; cooperative learning activities; and curricular development sessions. Students were involved in sessions on ethics, politics, economics and law. Teachers organize bioethics curriculum writing sessions, laboratory activities using electrophoresis as well as other biotechnology, and sessions on genetic databases.

A World Wide Web *home page* for Genetics Education assists teachers in remaining current on genetic information and helps them find answers to student inquiries. The *home page* has links to numerous genome sites, sources of information on genetic conditions, networking opportunities with other genetics education programs, teaching resources, lesson plan ideas, and the Mentor Network of genetic professionals and a network of family support groups willing to work with teachers and their students.

The home page is located at URL: *http://www.kumc.edu/GEC*

* *Supported by a grant from the U.S. Department of Energy, Human Genome Program under contract DE-FG02-92ER613.*

Human Genome Education Program

Lane Conn
Human Genome Education Program, Stanford Human Genome Center, Palo Alto, CA 94304
Telephone: (415) 812-2003; Fax: (415) 812-1916

The Human Genome Education Program (HGEP) operates within the Stanford Human Genome Center. It is a collaborative effort among HGEP staff, Genome Center scientists, collaborating staff from other education programs, experienced high school teachers, and an Advisory Panel in the fields of science, education, social science, assessment, and ethics.

The Human Genome Project will have a profound impact on society with its applications in testing for and improving treatment of genetic disease and the many uses of DNA profiling. The goal of HGEP is to help prepare high school students and community members to be able to make educated decisions on the personal, ethical, social and policy questions raised by the application of genome information and technology in their lives.

The primary objectives for HGEP are to (1) develop a human genome curriculum for high school science and (2) education outreach to schools and community groups in the San Francisco Bay Area. To achieve Objective 1, the HGEP is working to develop, field test, and prepare for national dissemination a two laboratory-based curriculum units for high school students. Unit 1, "Dealing With Genetic Disorders," explores the variety of treatment options potentially available for a genetic disorder, including gene therapy. Unit 2, "DNA Snapshots, Peeking at Your DNA," explores human relatedness through examining the student's own DNA polymorphisms using PCR.

Each unit is centered around a societal or ethical problem raised by these important applications of genome information and technology. Students use modeling exercises and inquiry laboratory experiments to learn about the science behind a given application. Students then combine the science they have learned with other relevant information to choose a solution to the societal/ethical problem posed in the unit. As a culminating activity, the students work in groups to present and defend their solution.

To achieve Objective 2, the HGEP provides Genome Center tours for teacher, student and community groups that involve pre-tour lectures; tour exploration of genome mapping, sequencing and informatics; and post-tour lecture and discussion on genome applications, and their social and ethical implications. Also, the education program continues to work to establish and sustain local science education partnerships among schools, industry, universities and national laboratories.

Genome Educators

Catherine Pinkas and Sylvia Spengler

Human Genome Program, Life Science Division, Lawrence Berkeley Laboratory, MS 1-459, Berkeley, CA 94720
EM: CIPinkas@lbl.gov, Sylviaj@ux5.lbl.gov

Sponsored by Lawrence Berkeley National Laboratory's Human Genome Program, Genome Educators is an informal network of educational professionals who have an active interest in all aspects of genetics research and education. This national group includes scientists, researchers, educational curriculum developers, ethicists, health professionals, high school teachers and instructors at college and graduate levels, and others in occupations affected by genetic research.

Genome Educators is a unique collaborative effort dedicated to sharing information and resources to further understanding of current advances in the field of genetics. Seminars, workshops and special events are sponsored at frequent intervals. Genome Educators maintains an active world wide web site (URL: http://www.lbl.gov/Education/Genome). This site contains a calendar of events, directory of participating genome educators, and information about educational resources and reference tools. Participating genome educators may publish articles and talks of interest at this site. In addition, a monitored discussion group is maintained to facilitate dialog and resource sharing among participants.

YOUR WORLD/OUR WORLD - BIOTECHNOLOGY & YOU SPECIAL ISSUE ON THE HUMAN GENOME PROJECT

Jeff Davidson (Executive Director), Laurence Weinberger, Esq. (Education Committee Co-Chair), Pennsylvania Biotechnology Association, State College, PA 16801*

Your World/Our World is a biotechnology science magazine published semi-annually by the non-profit Pennsylvania Biotechnology Association (PBA) describing for seventh to tenth grade students the excitement and achievements of contemporary biotechnology. This is the only continuing source of biotechnology education specifically directed to this age group - an age at which students too frequently are turned off from science. **The special Spring 1996 issue will be devoted to the presentation of the science behind the HGP, the HGP itself, and the ethical, legal, and social issues generated by the project.** The strong emphasis on attractive graphic presentation and age appropriate text that have been the hallmark of the earlier issues, which have been highly acclaimed and well received by the educational, scientific, and business community, will be continued.

PBA believes that increased educational opportunities to learn about biotechnology are most effective if presented at the seventh to tenth grade levels for the following reasons:

- Full semester life science and biology classes often occur for the first time in these grades;
- Across the nation, textbooks are typically 10 to 14 years old, and even the most recent textbooks are quickly dated by the rapid development in the biological sciences;
- Curricula at this level are more flexible than high school curricula, allowing the addition of information about exciting biological developments; and
- Science at this level is generally not elective, and, therefore, a very comprehensive student population is addressed rather than the more selective populations available later in the educational program.

In creating *Your World/Our World*, the PBA defined the following educational goals to guide the development of the magazine:

- Contribute to general science literacy and an educated electorate;
- Contribute to biological and technological literacy; and
- Motivate students to pursue additional science study and careers in science, particularly among women and minority populations.

PBA recognizes that it has been a point of pride that biotechnologists have been uniquely concerned with the impact of their technology on society and have been the first to raise and encourage responsible public debate without being forced to do so by others. To do less now for the children would be a breach of this responsible history. Accordingly, this special HGP issue will address the ethical, legal, and social issues raised by the new genomic technologies. Special ethics advisors have been recruited to aid in the development of these aspects.

A complimentary copy of the special issue and its teachers' guide will be **mailed to every public and private school seventh to tenth grade science teacher (approximately 40,000) in the United States.** A cover announcement will explain the origin and development of the magazine and of the special edition. Teachers will be invited to purchase full classroom packets (30 copies & teacher's guide) from the PBA, but, if they are not able to afford the packets, they will be asked to respond by postcard indicating their interest. The cost of the packets will probably be in the $20 range. The PBA is actively seeking additional support so that the issue may be distributed for free or at a reduced cost. In addition, parts of the special issue will be available over the Internet via a World Wide Web Page.

PBA believes this is a unique opportunity to educate America's youth about the HGP and insure that accurate non-sensational information will be made available to our country's children.

*Supported in part by a grant from the Office of Energy Research, U.S. Department of Energy.

NONTRADITIONAL INHERITANCE: GENETICS AND THE NATURE OF SCIENCE INSTRUCTIONAL MATERIALS FOR HIGH SCHOOL BIOLOGY

(U.S. Department of Energy Grant DE-FG03-95ER61989)

Joseph D. McInerney, B. Ellen Friedman, Biological Sciences Curriculum Study (BSCS), 5415 Mark Dabling Blvd., Colorado Springs, Colorado 80918

There often is a gap between the public's and scientists' views of new research findings, particularly if the public's understanding of the nature of science is not sound. Large quantities of new evidence and consequent changes in scientific explanations, such as those associated with the Human Genome Project and related genetics research, can accentuate those different views. Yet an appealing secondary effect of the unusually fast acquisition of data is that our view of genetics is changing rapidly during a brief time period, a relatively recent phenomenon in the field of biological sciences. This situation provides an outstanding opportunity to communicate the nature and methods of science to teachers and students, and indirectly to the public at large. The immediacy of new explanations of genetic mechanisms lets nontechnical audiences actually experience a changing view of various aspects of genetics, and in so doing, gain an appreciation of the nature of science that rarely is felt outside of the research laboratory.

BSCS is developing a curriculum module that brings this active view of the nature and methods of science into the classroom via examples from recent discoveries in genetics. We will distribute this print module free of charge to interested high school biology teachers in the United States.

The examples selected for classroom activities include the instability of trinucleotide repeats as an explanation of genetic anticipation in Huntington disease and myotonic dystrophy, and the more widespread genetic mechanism of extranuclear inheritance, illustrated by mitochondrial inheritance. Background materials for teachers discuss a wider range of phenomena that require nontraditional views of inheritance, including RNA editing, genomic imprinting, transposable elements, and uniparental disomy. The genetics topics in the module share the common characteristic that they are not adequately explained by the traditional, Mendelian concepts that are taught in introductory biology at the high school level. In addition to updating the genetics curriculum and communicating the nature of science, the module devotes one activity to the ethical and social aspects of new genetics discoveries by challenging students to consider the current reluctance to test asymptomatic minors for the presence of the HD gene.

The major challenge we have faced in this project is to make relatively technical genetics information accessible to high school teachers and students and to turn the often passive treatment of scientific processes into an active experience that helps students develop an understanding and appreciation of the nature and methods of science. The module is being field tested in classrooms across the country. Evaluation data from the field test will guide final revision of the module prior to distribution.

THE HUMAN GENOME PROJECT: BIOLOGY, COMPUTERS, AND PRIVACY
Development of Educational Materials for High School Biology

(U.S. Department of Energy Grant DE-FG03-93ER61584)

Joseph D. McInerney, Lynda B. Micikas, and *B. Ellen Friedman*, Biological Sciences Curriculum Study (BSCS), 5415 Mark Dabling Blvd., Colorado Springs, Colorado 80918

One of the challenges faced by the Human Genome Project (HGP) is to handle effectively the enormous quantities and types of data that emerge as a result of progress in the project. The informatics aspect of the HGP offers an excellent example of the interdependence of science and technology. In addition, the electronic storage of genomic information raises important questions of ethics and public policy, many revolving around privacy.

BSCS addresses the scientific, technological, ethical, and policy aspects of genome informatics in the instructional program titled *The Human Genome Project: Biology, Computers, and Privacy*. The program, intended for use in high school and college biology, consists of software and a 150-page print module. The software includes two model databases: a research database housing anonymous data (map data, sequence data, and biological/clinical information) and a registry that attaches names of 52 fictitious individuals (three kindreds) to genomic data. Students manipulate the database software as they work through seven classroom inquiries described in the print material. Also included is 50 pages of background material for teachers.

An introductory activity lets students become familiar with the software and dramatically demonstrates the advantages of technology in analysis of sequence data. In activities 1 and 2, students use the database to construct pedigrees and make initial choices about privacy with regard to genetic tests for their fictitious person. Activity 3 expands genetic anticipation, and in activities 4 and 5, students deal in depth with decision-making, ethics, and public policy, revisiting their earlier decision about testing and data accessibility. A final extension activity shows how comparisons with genomic data can be used to test hypotheses about the biological relationships between individual humans and about the evolutionary significance of DNA sequence similarities between different species.

External reviews and evaluation data from a field test involving 1,000 students in schools across the United States were used to guide final revision of the materials. BSCS will distribute the module free of charge to more than 10,000 high school and college biology teachers.

A HISPANIC EDUCATIONAL PROGRAM FOR SCIENTIFIC, ETHICAL, LEGAL AND SOCIAL ASPECTS OF THE HUMAN GENOME PROJECT*

Margaret C. Jefferson and Mary Ann Sesma, Department of Biology and Microbiology, California State University, Los Angeles CA 90032 and Los Angeles Unified School District.

The primary objectives of this grant are to develop, implement, and distribute culturally competent, linguistically appropriate, and relevant curriculum that leads to Hispanic student and family interactions regarding the science, ethical, legal, and social issues of the Human Genome Project. By opening up channels of familial dialogue between parents and their high school students, entire families can be exposed to genetic health and educational information and opportunities. In addition, greater interaction is anticipated between students and teachers, and parents and teachers. In the Los Angeles Unified School District alone, over 65% of the approximately 850,000 student enrollment are bilingual Hispanics. The 1990 census data revealed that the U.S.A. had a total population of 248,709,873, of which 22,354,059 were Hispanics, and thus, there is a need for materials to be disseminated throughout the U.S.A. that are relevant and understandable to this population.

Student curriculum consists of BSCS HGP-ELSI curriculum available in both English and Spanish; supplemental lesson plans developed and utilized by high school teachers in predominantly Hispanic classrooms that will be available via the World Wide Web; student-developed surveys that ascertain knowledge and perceptions of genetics and HGP-ELSI in Hispanic and other ethnic communities in the greater Los Angeles area; the University of Washington High School Human Genome Program exercises on DNA synthesis and sequencing; and career ladders and opportunities in genetics. The supplemental lesson plans are focused on four major units: the Cell; Mendelian Genetics and its Extensions; Molecular Genetics; and the Human Genome Project and ELSI. The concise concepts underlying each unit are being utilized in two ways: (a) first, the student activities emphasize logical, problem-solving exercises; tools or technologies applicable to that concept; when and where appropriate, a focus on the Hispanic population; and an understanding of the problems and compassion for the families associated with learning of genetic diseases. (b) second, the concepts serve as the springboard for the topics that the students include in science newsletters to their parents. In addition to on-campus activities, we intend to arrange field trips and/or classroom demonstrations of genetic and molecular biology techniques by scientists and other experts. The speakers would also be asked to discuss career opportunities and the educational requirements needed to enter the specific careers presented.

The parent curriculum consists of two major activities. First the student-parent newsletter is designed to drawn the parents into the curriculum. Students write newsletters on a biweekly basis. Each newsletter relates to a student curriculum subunit and the specific subunit concepts. English, Spanish, social science as well as biology and chemistry teachers assist the students in its production. The other major activity that involves the parents are the parent focus groups. Parents from each participating school are invited to monthly focus groups at their specific campus. The focus groups discuss issues related to genetics and health, legal and social issues as well as science issues that stem from the student newsletters. The discussions are in both English and Spanish with translators available. Links with other programs have been established.

*Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy under grant # DE-FG03-94-ER61797.

Involvement of High School Students in the Sequencing of the Human Genome

Maureen M. Munn, Maynard V. Olson and Leroy Hood

Department of Molecular Biotechnology
University of Washington, Seattle, WA 98195

For the past two years, we have been developing a program that involves high school students in the excitement of genetic research by enabling them to participate in sequencing the human genome. This program provides high school teachers with the proper training, equipment, and support to lead their students through the exercise of sequencing small portions of DNA. The participating classrooms carry out two experimental modules, DNA synthesis (an introduction to DNA replication and the techniques used to study it) and DNA sequencing. Both of these experiments consist of three parts- synthesizing DNA fragments using Sequenase and a biotin-labeled primer, bench top electrophoresis using denaturing polyacrylamide gels, and colorimetric DNA detection that is specific for the biotinylated primer. Students analyze their sequencing data and enter it into a DNA assembly program. This year, in collaboration Eric Lynch and Mary-Claire King from the Department of Genetics at the University of Washington, the students will be sequencing a region of chromosome 5q that may be involved in a form of hereditary deafness.

Students also consider the ethical, legal and social issues (ELSI) of genome research in a unit that explores the topic of presymptomatic testing for Huntington's disease (HD). This module was developed by Sharon Durfy and Robert Hansen from the Department of Medical History and Ethics at the University of Washington. It provides a scenario about a family that carries the HD allele, descriptions of the clinical and genetic aspects of the disorder, an exercise in drawing pedigrees and an autoradiograph showing the PCR assay used to detect HD. Students use an ethical decision-making model to decide whether, as a character from the scenario, they would be tested presymptomatically for the HD allele. Through this experience, they develop the skills to define ethical issues, ask and research the relevant questions about a particular topic and make justifiable ethical decisions.

In the first two years of this program, our focus was on the development of robust, classroom-friendly modules that can be presented in up to six classes at one time. This year we will focus on disseminating this program to local, regional, and national sites. During a week-long workshop in July, 1995, we trained an additional thirteen high school teachers, bringing our current number to twenty teachers at thirteen schools. We have recruited local scientists to act as mentors to each of the schools and provide classroom support. On the regional level, four of our teachers are from outside the greater Seattle area and will be supported during the classroom experiments by scientists in their region. We have presented this program at national meetings and workshops, including the Human Genome Teacher Networking Project Workshop in Kansas City, KS (June, 1995) and the meeting of the National Association of Biology Teachers in Phoenix, AZ (October, 1995). We have also distributed our modules to teachers and scientists throughout the nation to encourage the development of similar programs. This year we will also develop and pilot a module using automated sequencing. This will enable distant schools to participate in the program by providing them with the option of sending their DNA samples to the UW genome center for electrophoresis.

While we hope the human genome sequencing experience will interest some students in science careers, a broader goal is to encourage high school students to think constructively and creatively about the implications of scientific findings so that the coming generation of adults will make judicious decisions affecting public policies.

This work is sponsored by the U.S. Department of Energy under grant No. DE-FG06-94ER61798.

COMMUNITY COLLEGE INITIATIVE
A Program for Preparing Community College Students for Work in Biotechnology

Catherine Pinkas, Laurel Egenberger*, Life Sciences Division, *Center for Science and Engineering Education. Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, Ca. 94720

The Community College Initiative is a program to prepare community college students for work in biotechnology. This program is a combined effort of Lawrence Berkeley National Laboratory (LBNL) and the California Community Colleges. Its goal is to develop mechanisms to encourage students, particularly from underrepresented groups, to pursue science, mathematics, engineering and technology studies, to participate in forefront laboratory research, and to gain valuable work experience.

The initiative is structured to upgrade the skills of students and their instructors through four major program components spread over three years.

SUMMER STUDENT WORKSHOPS: Four week summer residential programs for students who have completed the first year of their biotechnology academic studies. Students attend an orientation program and then work through a series of realistic laboratory exercises to develop their problem solving, decision making, laboratory techniques, computer, instrumentation, and mathematics skills. Ethical, legal, and social concerns are integrated into the program and students learn to identify commonly shared values of the scientific community and increase their understanding of issues of personal and public concern.

TEACHER WORKSHOP TRAINING: Seminars for biotechnology instructors to improve, upgrade, and update their understanding of current technology and laboratory practices. Special emphasis is placed on laboratory techniques, instrumentation, computers (including the Internet), safety, and curriculum development in current topics in ethical, legal, and social issues in science.

SABBATICAL FELLOWSHIPS: Semester length research fellowships for community college instructors, provides investigative and field experience in research laboratories. Instructors will work one-on-one with research scientists on specific prearranged projects. These projects may vary from traditional research to creation of training materials for use in the laboratory and/or community college classroom.. During the fellowship, these teachers also assist in development of the second year student summer research activities for their project.

SUMMER FACULTY-STUDENT TEAMS: Faculty who have completed research fellowships and students who have finished their second academic year join forces on projects begun in the teacher sabbatical program. The community college instructors provide the much needed assistance in developing curriculum, providing laboratory training, and assistance mentoring summer students during this ten week program.

IMPLICATIONS OF THE GENETICIZATION OF HEALTH CARE FOR PRIMARY CARE PRACTITIONERS*

Mary B. Mahowald, John Lantos, Mira Lessick, Robert Moss, Lainie Friedman Ross, Greg Sachs, Marion Verp, Department of Obstetrics and Gynecology and MacLean Center for Clinical Medical Ethics, University of Chicago, Chicago, IL 60637 (http://ccme-mac4.bsd.uchicago.edu/CCMEHomePage.html)

"Geneticization" refers to the process by which advances in genetic research are increasingly applicable to all areas of health care.[1] Studies show that primary caregivers are often deficient in their knowledge of genetics and genetic tests, and the ethical, legal, and social implications of this knowledge.[2-6] Accordingly, this project prepares primary caregivers who have no special training in genetics or genetic counseling to deal with the implications of the Human Genome Project for their practice.

Phase I (fall 1995): Generic topics will be addressed by PI and Co-PIs with Robert Wood Johnson clinical scholars and clinical ethics fellows, led by visiting or internal experts.
Topics: Goals, Methods, & Achievements of the HGP; Typology of Genetic Conditions; Scientific, Clinical, Ethical, and Legal Aspects of Gene Therapy; Concepts of Disease; Genetic Disabilities; Gender and Socio-economic Differences; Cultural and Ethnic Differences; Directive or Nondirective genetic counseling.
Speakers: Jeff Leiden; Julie Palmer; Dan Brock; Anita Silvers; Abby Lippman; James Bowman; Beth Fine

Phase II (Jan.-Mar., 1996): Teams of individuals, all trained in the same area of primary care, will identify & address issues specific to their area, developing course outlines, bibliography, and methodology based on grand rounds given by national expert.

Primary care area	Genetics expert	Ethics expert	
Pediatrics	Stephen Friend	Lainie Friedman Ross	+ fellow
Obstetrics/Gynecology	Joe Leigh Simpson	Marion Verp	+ fellow
Medicine	Tom Caskey	Greg Sachs	+ fellow
Family medicine	Noralane Lindor	Robert Moss	+ fellow
Nursing	Mira Lessick	Colleen Scanlon	+ fellow

Phase III (Apr,-May, 1996): Policy issues will be identified and addressed as above for all areas of primary care, based on grand rounds given by national expert.
Policy team: Genetics expert: Sherman Elias; Ethics expert: John Lantos + trainee

Phase IV (Oct.-Dec. 1996): Presentation of content developed to new group of fellows and scholars by each of the above teams, followed by evaluation & revision.

Phase V (spring, 1997): NATIONAL CONFERENCE and CME/CNE WORKSHOPS for primary caregivers, keynoted by Victor McKusick.

*Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy, under contract DE-FG02-95ER61990.

[1]Lippman A., Prenatal genetic testing and screening, *Amer J Law & Med* XVII, 15-50 (1991).
[2]Hofman, K.J., Tambor, E.S., Chase, G.A., Geller, G., Faden, R.R., and Holtzman, N.A., Physicians' knowledge of genetics and genetic tests, *Acad Med* 68, 625-32 (1993).
[3]Holtzman, N.A., The paradoxical effect of medical training, *J Clin Ethics* 2, 241-42 (1992).
[4]Forsman, I, Education of nurses in genetics, *Amer J of Hum Genetics* 552-58, (1988).
[5]Williams, J.D., Pediatric nurse practitioners' knowledge of genetic disease *Ped Nursing* 9, 119-21 (1983).
[6]George, J.B., Genetics: Challenges for nursing education, *J Ped Nursing* 7 , 5-8, (1992).

GETTING THE WORD OUT ON THE HUMAN GENOME PROJECT: A COURSE FOR PHYSICIANS

Sara L. Tobin, Department of Biochemistry and Molecular Biology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104 and *Ann Boughton*, Thumbnail Graphics, 228 Northwest 32nd Street, Oklahoma City, OK 73118.

Progressive identification of new genes and implications for medical treatment of genetic diseases appear almost daily in the scientific and medical literature, as well as in public media reports. However, most individuals do not understand the power or the promise of the current explosion in knowledge of the human genome. This is also true of physicians, most of whom completed their medical training prior to the application of recombinant DNA technology to medical diagnosis and treatment. This lack of training prevents physicians from appreciating many of the recent advances in molecular genetics and may delay their acceptance of new treatment regimens. In particular, physicians practicing in rural communities are often limited in their access to resources that would bring them into the mainstream of current molecular developments. This project is designed to fill two important functions: first, to provide solid training for physicians in the field of molecular medical genetics, including the impact, implications, and potential of this field for the treatment of human disease; second, to utilize physicians as informed community resources who can educate both their patients and community groups about the new genetics.

We propose to develop a flexible, user-friendly, interactive multimedia CD-ROM designed for continuing education of physicians in applications of molecular medical genetics. To initiate these objectives, we will develop the design of the CD and will produce a prototype providing a detailed presentation of one of the four training areas. These areas are (1) Genetics, including DNA as a molecular blueprint, chromosomes as vehicles for genetic information, and patterns of inheritance; (2) Recombinant techniques, stressing cloning and analytical tools and techniques applied to medical case studies; (3) Current and future clinical applications, encompassing the human genome project, technical advances, and disease diagnosis and prognosis; and (4) Societal implications, focusing on approaches to patient counseling, genetic dilemmas faced by patients and practitioners, and societal values and development of an ethical consensus. Area (2) will be presented in the prototype.

The CD format will permit the use of animation, video, and audio, in addition to graphic illustrations and photographs. We will build on our existing base of computer generated illustrations. A hypertext glossary, user notes, practice tests, and customized settings will be utilized to tailor the CD to the needs of the user. Brief, multiple-choice examinations will be evaluated for continuing medical education credits by the Office of Continuing Medical Education. The CD will be programmed to permit updates of scientific and medical advances either by downloading from the Internet or from a disc available by subscription.

This is a cooperative project involving individuals with documented expertise in teaching of molecular medical genetics, continuing medical education, graphic design, and CD-ROM production. The content of the CD will be supervised by a scientific board of directors. We present mechanisms for the evaluation of the CD by rural Oklahoma physicians. Arrangements have been made for distribution of the CD by a national publisher of medical and scientific materials. This CD will provide a powerful tool to educate physicians and the public about the power and potential of the human genome project for the benefit of human health.

AAAS Congressional Fellowship Program

Stephen Goodman
The American Society of Human Genetics; Bethesda, MD 20814-3998
301/571-1825, Fax: /530-7079, *society@genetics.faseb.org*

Few individuals in the genetics community are conversant with federal mechanisms for developing and implementing policy on human genetics research. In 1995 the American Society of Human Genetics (ASHG), in conjunction with DOE, initiated an American Association for the Advancement of Science (AAAS) Congressional Fellowship Program to strengthen the dialogue between the professional genetics community and federal policymakers. The fellowship will allow genetics professionals to spend a year as special legislative assistants on the staff of members of Congress or on congressional committees. Directed toward productive scientists, the program is intended to attract independent investigators.

In addition to educating the scientific community about the public policy process, the fellowship is expected to demonstrate the value of science-government interactions and make practical contributions to the effective use of scientific and technical knowledge in government. The program includes an orientation to legislative and executive operations and a year-long weekly seminar on issues involving science and public policy.

Unlike similar government programs, this fellowship is aimed primarily at scientists outside government. It emphasizes policy-oriented public service rather than observational learning and designates its fellows as free agents rather than representatives of their sponsoring societies.

One of the goals of DOE and ASHG is to develop a group of nongovernmental professionals who will be equipped to deal with issues concerning human genetics policy development and implementation, particularly in the current environment of health-care reform and managed care. Graduates of this program will serve as a resource for consultation in the development of public-health policy concerning genetic disease.

Fellowship candidates must demonstrate exceptional basic understanding of and competence in human genetics; hold an earned degree in genetics, biology, life sciences, or a similar field; have a well-grounded and appropriately documented scientific and technical background; have a broad professional background in the practice of human genetics as demonstrated by national or international reputation; be cognizant of related nonscientific matters that impact on human genetics; exhibit sensitivity toward political and social issues; have a strong interest and some experience in applying personal knowledge toward the solution of social problems; be a member of ASHG; be articulate, literate, adaptable, and interested in working on long-range public policy problems; be able to work with a variety of people of diverse professional backgrounds; and function well during periods of intense pressure.

The first fellow is working in the office of Senator Wellstone, Democrat from Minnesota, and devoting most of his time to studying and commenting on health-care and science issues.

Pathways to Genetic Screening: Molecular Genetics Meets the High Risk Family

By Troy Duster, et al.

Institute for the Study of Social Change
University of California, Berkeley, CA 94705

The proliferation of genetic screening and testing is requiring increasing numbers of Americans to integrate genetic knowledge and interventions into their family life and personal experience. This study examines the social processes that occur as families at risk for two of the most common autosomal recessive diseases, sickle cell disease (SC) and cystic fibrosis (CF), encounter genetic testing. Since each of these diseases is found primarily in a different ethnic/racial group (CF in European Americans and SC is African Americans), this research will clarify the role of culture in integrating genetic testing into family life and reproductive planning. A third type of genetic disorder, the thalassemias, has recently been added to our sample in order to extend our comparative frame to include other ethnic and racial groups. In California, the thalassemias primarily affect Southeast Asian immigrants, although another risk group is from the Mediterranean region. Thalassemias, like cystic fibrosis and sickle cell disease, have a similar pattern of inheritance and raise similarly serious bio-medical challenges and issues of information management.

Data are drawn from interviews with members of families in which a gene for CF, SC or thalassemia has been identified. Data collection consists primarily of focused interviews with approximately 400 individuals from families in which at least one member has been identified as having a genetic disorder (or trait). In the most recent phase of the research, we are conducting focus groups selected to achieve stratified homogeneity around key social dimensions such as gender and relationship to disease. This is clarifying the social processes that facilitate and inhibit genetic testing.

We are currently assessing the concerns expressed by respondents about the potential uses of genetic information. We find strong patterns of concern, often based on personal experience, that genetic information may be used in ways that family members perceive as dangerous and/or discriminatory. First among these concerns is fear of losing access to health care. Additional concerns include fear of genetic discrimination in employment and other types of insurance, particularly life insurance. Similar patterns of concern exist among members of each ethnic group, and are frequently the focus of attention among family members, but take somewhat different form within each cultural group. These concerns constitute a growing obstacle to widespread use of genetic testing.

This work was supported by the Director, Office of Energy Research, Office of Health and Environmental Research of the United States Department of Energy under contract DE-FG03-92ER61393.

THE GENETICS ADJUDICATION RESOURCE PROJECT

The Einstein Institute for Science, Health & the Courts is preparing the foundation for a new utility needed to prepare the nation's 21,000 courts to adjudicate the genetics and ELSI-related issues that foreseeably will rush into the courtroom as the Human Genome Project completes its genomic mapping and sequencing mission during the next ten years. This project initiates practical collaboration among courts, legal and policy-making institutions, and science centers leading to modalities for understanding the scientific validity of claims, and for the resolution of ethical, legal, and social disputes arising within the genetic testing and gene therapy contexts. Our objective over the ensuing decade is to facilitate genetic testing and gene therapy dispute management, and to avoid to the extent possible the confusion that characterized adjudication of forensic DNA technologies during the decade just ended.

The outlines of a genetics adjudication utility were given form by the 1995 Working Conversation on Genetics, Evolution, and the Courts, involving 37 federal and state judges in others in science and policy-making leadership positions from across the nation. The courts are becoming aware of genetics, molecular biology, and their applications, and judges want public confidence to be maintained as the profound and complex issues set in motion by the HGP begin the long course of litigation. Modalities for understanding the underpinning science are needed, as well as instrumentalities to assure that the best cases are actually filed and pursued. Because the courts are the front-line for resolving disputes, creative lawyering will assure an abundance of lawsuits. Many such lawsuits will request the courts to make policy judgments, perhaps best undertaken by state legislatures and Congress. Accordingly, a new adjudication utility should provide forums for judicial/legislative exchange, preparatory deliberations in anticipation of pressure to make rushed policies under conditions of great social uncertainty in the wake of human genetics progress.

EINSHAC will provide a design, planning, communications, and implementation center for a multi-purpose resource project available to the courts. It will undertake over an 18 month period the following tasks, pilot-testing each and assessing the best organizational locales for those that exhibit promise:

1. Judicial Education in Genetics & ELSI-Related Issues for six Judicial Branch leadership associations and nine metropolitan courts -- aimed at 1,000 judges -- in conjunction with scientific faculty and coaches mobilized by DOE/national laboratories and the American Society for Human Genetics.
2. Judicial Digital Electronic Collegium -- technological modernization of the courts community by providing access to ELSI and genetics information through Internet resources.
3. Amicus Brief Development Trust Fund -- a process and resources to support law development at the state and federal appeals courts level.
4. Genetics Indigent Party Trust Fund -- a process and resources at the state and federal trial level to sustain meritorious civil cases holding promise of effective law development.
5. Establishment of a Pro-Bono Legal Services Clearinghouse -- a personal and on-line referral resource for persons seeking representation for genetics and ELSI-related cases.
6. Access to Neutral Expert Witnesses -- advisors to courts encountering particularly complex cases deemed right for the judicial exercise of Federal Rule of Evidence 706 and its State counterparts.
7. Pilot of Judicial/Legislative ELSI Policy Forums -- provision of neutral staff and coordination in three mid-Atlantic states considering legislation related to health care, insurance, privacy, medical records.
8. National Training Center for Minority Justice Personnel -- facilitating a leadership preparation program for the nation's minority court-related personnel in a consortium arrangement with the Ruffin Society of Massachusetts, the College of Criminal Justice at Northeastern University, and the Flaschner Judicial Institute.

The Project actively involves judges, scientists, and prominent lawyers. It will report to the *EINSHAC* Board of Directors that includes prominent judges, justices and scientists, several of whom participated in the 1995 Working Conversation on Genetics, Evolution and the Courts. As a continuing guidance forum, *EINSHAC* will conduct a Working Conversation followup in Orleans, Cape Cod in July, 1996.

For additional information, please contact Dr. Franklin M. Zweig, President, Einstein Institute for Science, Health and the Courts, Suite 750, 3 Bethesda Metro Center, Bethesda, Maryland 20814, Tel (301) 961-1949 - Fax (301) 913-0448 - E-Mail Einshac@aol.com

INTELLECTUAL PROPERTY ISSUES IN GENOMICS*

Rebecca S. Eisenberg, University of Michigan Law School, Ann Arbor, MI 48109

Intellectual property issues have been uncommonly salient in the recent history of advances in genomics. Beginning with the filing of patent applications by NIH on the first batch of expressed sequence tags (ESTs) from the laboratory of Dr. Craig Venter, each new development has been met with speculation about its strategic significance from an intellectual property perspective. Are ESTs of unknown function patentable, or is further work necessary before they satisfy patent law standards? Will patents on such fragments promote commercial investment in product development, or will they interfere with scientific communication and collaboration and retard the overall research effort? Without patent rights, how may the owners of private cDNA sequence databases earn a return on their investment while still permitting other investigators to obtain access to the information on reasonable terms? What are the rights of those who contribute resources such as cDNA libraries that are used to create the databases, and of those who identify sequences of interest out of the morass of information in the databases by formulating appropriate queries? Will the disclosure of ESTs in the public domain preclude patenting of subsequently characterized full-length genes and gene products? And why would a commercial firm invest its own resources in generating an EST database for the public domain?

Two factors have contributed to the fascination with intellectual property in this setting. First is a perception that some pioneers in genomics have sought to claim intellectual property rights that reach beyond their actual achievements to cover future discoveries yet to be made by others. For example, the controversial NIH patent applications claimed rights not only in the ESTs that were actually set forth in the specifications, but also in the full-length cDNAs that might be obtained by using the ESTs as probes, as well as in other, undisclosed fragments of those genes. More recently, private owners of cDNA sequence databases have set as a condition for access agreement to offer the database owners licenses to any resulting intellectual property. These efforts to claim rights to the future discoveries of others raise issues about the fairness and efficiency of the law in allocating rewards and incentives along the path of cumulative innovation.

Second is the counterintuitive alignment of interests in the debate. It was a public institution, NIH, that initially favored patenting discoveries that some representatives of industry thought should remain unpatented, and it was a major pharmaceutical firm, Merck & Co., that ultimately took upon itself the quasi-governmental function of sponsoring a university-based effort to place comparable information in the public domain. These topsy-turvy positions in the public and private sectors raise intriguing questions about the proper roles of government and industry in genomics research, and about who stands to benefit (and who stands to lose) from the private appropriation of genomic information.

---

*Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy under contract DE-FG02-94ER61792.

Infrastructure

This page intentionally left blank.

Human Genome Task Group

Benjamin J. Barnhart, Daniel W. Drell, Marvin Frazier, Gerald Goldstein, Jay Grimes, Roland Hirsch, Robert J. Robbins[1], Jay Snoddy, Marvin Stodolsky, David G. Thomassen, and John C. Wooley.
Chair: David A. Smith, 1993-1996; Aristides Patrinos, effective February 1996

Office of Health and Environmental Research, Health Effects and Life Sciences Division, Department of Energy, 19901 Germantown Road, Germantown, MD 20874-1290; (301)903-6488; FAX (301)903-8521; Internet *genome@er.doe.gov.*

Human Genome Coordinating Committee

Elbert Branscomb, Anthony Carrano, C. Thomas Caskey,[2] Chris Fields, Raymond Gesteland, Leroy Hood, David Kingsbury, Robert Moyzis, Mohandas Narla, Lloyd Smith, Mike Palazzola, and Lisa Stubbs.
Chair: David A. Smith, 1993-1996; Aristides Patrinos, effective February 1996
Executive Officer: Sylvia Spengler

Health and Environmental Research Advisory Committee

Keith O. Hodgson **(Chair),** Eugene W. Bierly, Mina J. Bissell, E. Morton Bradbury, Rita R. Colwell, Hadi Dowlatabadi, Jonathan Greer, Leroy E. Hood, Fern Y. Hunt, David T. Kingsbury, Gordon J.F. MacDonald, Jerold Melillo, Jill P. Mesirov, James W. Mitchell, Melvin I. Simon, Janet L. Smith, Henry N. Wagner, Jr., Susan S. Wallace, Warren M. Washington, Sheldon Wolff, James H. Wyche, and W. Franklin Harris, III.

HERAC may occasionally form subcommittees with HERAC members and other experts to help HERAC provide advice in specific research fields. Relevant current HERAC subcommittees include an instrumentation subcommitee (Keith O. Hodgson, Chair) and a human genome subcommittee (Melvin I. Simon, Chair).

---

[1] Now at Fred Hutchinson Cancer Research Center; Seattle, WA 98104

[2] Through 1995

This page intentionally left blank

Human Genome Management Information System

Betty K. Mansfield, Anne E. Adamson, Denise K. Casey, Sheryl A. Martin, **John S. Wassom**, Judy M. Wyrick, Laura N. Yust, Murray Browne, and Marissa D. Mills

Biomedical and Environmental Information Analysis Section; Health Sciences Research Division; Oak Ridge National Laboratory; 1060 Commerce Park, MS 6480; Oak Ridge, TN 37830
423/576-6669, Fax: /574-9888, *bkq@ornl.gov*

The Human Genome Management Information System (HGMIS), which was inaugurated in 1989, provides technical communication and information services for the Human Genome Program Task Group of the DOE Office of Health and Environmental Research. HGMIS facilitates research by (1) helping to communicate genome-related matters and investigations to contractors, grantees, and other publications via hard copy and World Wide Web; (2) serving as a clearinghouse for information on the U.S. genome project; and (3) reducing duplication of research efforts by providing a forum for information exchange among Human Genome Project investigators worldwide. HGMIS also occasionally compiles and organizes administrative data for DOE.

Communication Through Hard-Copy Publications and WWW. To fulfill its communication goals, HGMIS publishes the bimonthly newsletter *Human Genome News (HGN)*. HGMIS also produces a primer on molecular genetics and reports on the DOE Human Genome Program, Santa Fe contractor-grantee workshops, and other related subjects; and makes many of its publications available via WWW (*http://www.ornl.gov/TechResources/Human_Genome/home.html*) and Gopher (*gopher.gdb.org*).

HGN features technical and general interest articles; meeting reports; national and international project news; articles on ethical, legal, and social implications of the genome project; features on progress, goals, informatics, mapping, sequencing, technology development and transfer, and resources for facilitating research; genome meeting and training calendars; and grant and fellowship announcements. Some 11,500 domestic and foreign subscribers include genome and basic researchers at universities, national laboratories, and other research institutions; professors and teachers; industry representatives; legal personnel; ethicists; students; genetic counselors; physicians; science writers; and other interested individuals. *HGN* also serves as a primary source for discipline-specific publications that extract or reprint information on the Human Genome Project. HGMIS staff members continuously monitor changes in direction of the international Human Genome Project and search for ways to strengthen the content relevancy of the newsletter and other HGMIS services.

HGMIS maintains and updates the Genetics section of the Virtual Library from CERN (Switzerland) and the DOE genome project–specific pages. HGMIS also collaborates with the Einstein Institute for Science, Health, and the Courts to help educate the judiciary on genetics and other biomedical issues via the WWW site. Between January and July 1995, usage of HGMIS WWW text files increased more than 20-fold, from 444 to 9326 information requests a month.

Document Distribution. HGMIS staff has distributed about 53,000 items requested by subscribers and meeting attendees, including program and workshop reports, the DOE-NIH 5-year plan, primer, and the DOE informatics summit report. Numerous copies of full or partial documents also have been distributed for educational purposes.

Information Requests. HGMIS staff members answer questions and supply general information about the Human Genome Project by telephone, fax, and e-mail. For example, those in biotechnology and other industries use HGMIS as a resource for identifying goods and services that might be useful to genome researchers. HGMIS also links callers with appropriate Human Genome Project contacts with experience in the caller's subject area. HGMIS staff exchange ideas and suggestions with investigators, industry representatives, and others when they display the DOE Human Genome Project traveling exhibit at scientific conferences and genome-related meetings.

HGMIS invites comments and suggestions about its documents and services, which are available upon request and without charge.

This work is sponsored by the Office of Health and Environmental Research, U.S. Department of Energy, under contract No. DE-AC05-84OR21400 with Lockheed Martin Energy Systems, Inc.

HUMAN GENOME PROGRAM COORDINATION

Sylvia Spengler, Catherine Pinkas, Katalin Markowitz, Life Sciences Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720

The Human Genome Program Coordination office assists the management of the DOE Human Genome Program coordinating DOE-funded research and evaluating the progress and impact of the Program, as well as representing the Program when requested.

The Human Genome Coordinating Committee, HGCC, was created by the Office of Health and Energy Research in October 1988. Its members represent DOE Human Genome Program efforts at the national laboratory Genome Centers, as well as grantees of the program. Representatives of other agencies and groups, including NSF, HUGO, NCHGR, and USDA, are invited to attend HGCC meetings. In addition, the memorandum of understanding between DOE and NIH (October 1988) established a joint subcommittee of HERAC and the NIH Program Advisory Committee (now the Genome Council) to coordinate the research effort of the two US agencies most deeply involved in the genome program.

Field Operations is responsible for not only the activities of the HGCC and its subcommittees, but also for coordinating and organizing major meetings of the Programs, the joint DOE-NIH working groups, and providing scientific support for the development of new initiatives. The office also coordinates and participates in meetings of the task forces and working groups established by the HGCC in order to provide timely reports on activities to the HGCC and the Genome Program Manager. Other activities include writing and editing reports of committees and preparing materials for publication from the HGCC.

ALEXANDER HOLLAENDER DISTINGUISHED POSTDOCTORAL FELLOWSHIPS*

Linda Holmes and Eugene Spejewski, Oak Ridge Institute for Science and Education, P.O. Box 117, Oak Ridge, TN 37831-0117

The Alexander Hollaender Distinguished Postdoctoral Fellowships, sponsored by the Department of Energy (DOE), Office of Health and Environmental Research (OHER), support research in the fields of life, biomedical, and environmental sciences. Since the DOE Human Genome Distinguished Postdoctoral Fellowships and DOE Global Change Distinguished Postdoctoral Fellowships both had their last application cycles in FY 1995, the Hollaender program is now open to recent PhD graduates in the fields of human genome and global change, as well.

Fellowships of up to two years are tenable at any DOE, university, or private laboratory providing the proposed adviser at that laboratory receives at least $150,000 per year in support from OHER. Fellows earn stipends of $37,500 the first year and $40,500 the second. To be eligible, applicants must be U.S. citizens or permanent residents at the time of application, and must have received their doctoral degrees within two years of the earliest possible starting date, which is May 1 of the appointment year.

The Oak Ridge Institute for Science and Education (ORISE), administrator of the fellowships, prepares and distributes program literature to universities and laboratories across the country, accepts applications, convenes a panel to make award recommendations, and issues stipend checks to fellows. The review panel identifies finalists from which DOE selects the award winners. Deadline for the FY 1997 fellowship cycle is January 15, 1997. For more information or an application packet, contact Linda Holmes at the Oak Ridge Institute for Science and Education, P.O. Box 117, Oak Ridge, TN 37831-0117 (423/576-9975, Fax: /241-5219).

*Administered by Oak Ridge Institute for Science and Education. ORISE and its programs are operated by Oak Ridge Associated Universities through a management and operating contract (DE-AC05-76OR00033) with the U.S. Department of Energy.

This page intentionally left blank.

Appendices

This page intentionally left blank.

Appendix A: Author Index

First authors are in **bold**.

C

D

E

F

G

H

I-J

K

L

M

N

O

P

Q-R

S

T

U-V

W

X-Y-Z

Appendix B: National Laboratory Index

U.S. Department of Energy Laboratories

Human Genome Program work at the national laboratories is described in the following abstracts.

This page intentionally left blank.