Advanced Detectors for Mass Spectrometry

W.H. Benner and **J.M. Jaklevic**
Human Genome Group; Engineering Science Department; Lawrence Berkeley National Laboratory; University of California; Berkeley, CA 94720
510/486-7194, Fax: -5857, *whbenner@lbl.gov*
http://www-hgc.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. We are also attempting to implement a rapid mass spectrometry-based screening procedure for PCR products.

The implementation of these applications requires that the performance of matrix-assisted-laser-desorption-ionization (MALDI) and electrospray mass spectrometry is improved. Our focus is the development of new ion detectors which will advance the state-of-the-art of each of these two types of spectrometers. 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 a type of cryogenic detector called a superconducting tunnel junction device which responds to the phonons produced when ions strike the detector. This detector does not rely on the formation of secondary electrons. We have demonstrated this type of detector to be at least two orders of magnitude more sensitive, on an area-normalized basis, than microchannel plate ion detectors. This development could extend the upper mass limit of MALDI-TOF-MS and increase sensitivity.

Electrospray ion sources generate ions of mega-Dalton 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 was 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. We have been able to mass analyze DNA molecules in the 1 to 10 MDa range using charge-detection mass spectrometry. In this technique, individual electrospray ions are directed to fly through a metal tube which detects their image charge. Simultaneous measurement of their velocity provides a way to measure their mass when ions of known energy are sampled. Several thousand ions can be analyzed in a few minutes, thus generating statistically significant mass values regarding the ions in a sample population. We are attempting to apply this technology to the analysis of PCR products.

DOE Contract No. DE-AC03-76SF00098.

Mass Spectrometer for Human Genome Sequencing

Chung-Hsuan Chen, Steve L. Allman, and K. Bruce Jacobson
Oak Ridge National Laboratory; Oak Ridge, TN 37831
423/574-5895, Fax: -2115, *chenc@ornl.gov*

The objective of this program is to develop an innovative fast DNA sequencing technology for the Human Genome Project. It can also be applied to fast screening of genetic and contagious diseases, DNA fingerprinting, and environmental impact analysis.

The approach of this program is to replace conventional gel electrophoresis sequencing methods by using lasers and mass spectrometry for sequencing. The present gel sequencing method usually takes hours to days to acquire DNA analysis or sequencing, since different lengths of DNA segments need to be separated in dense gel. With laser desorption mass spectrometry (LDMS) approach, various sizes of DNA segments are separated in the vacuum chamber of a mass spectrometer. Thus, the time taken to separate various sizes of DNA is less than one second compared to hours using other methods.

Recently, we successfully demonstrated sequencing short DNA segments with this approach. We also have succeeded in using LDMS for fast screening of cystic fibrosis disease. We succeeded in identifying both point mutation and deletion of cystic fibrosis. In addition, we had preliminary success in using LDMS to achieve DNA fingerprinting. Thus, laser desorption mass spectrometry (LDMS) is going to emerge as a new and important biotechnological tool for DNA analysis.

DOE Contract No. DE-AC05-84OR21400.

*Projects designated by an asterisk received small emergency grants following December 1992 site reviews by David Galas (formerly DOE Office of Health and Environmental Research, which was renamed Office of Biological and Environmental Research in 1997), Raymond Gesteland (University of Utah), and Elbert Branscomb (Lawrence Livermore National Laboratory).

Genomic Sequence Comparisons

George Church
Harvard Medical School; Boston, MA 02115
617/432-0503 or -7562, Fax: -7266
http://arep.med.harvard.edu

The first objective of this project is completion of an automated system to sequence DNA using electrophore mass-tag (EMT) primers for dideoxy sequencing. The prototype machine will contain a 60 capillary array with 400 EMT-labeled sequence ladders per capillary. The system is designed to use 100-fold less reagent and have 500-fold higher speed (1000 bases per sec per instrument) than current sequencing technology. Cleavage and laser desorption of EMTs from membranes for subsequent detection by EC-TOF mass spectrometry. The second objective is to overcome the limitations of purely hypothetical annotation of the growing number of reading frames in new genome sequences. We measure gene product levels and interactions using DNA microarrays, whole genome in vivo footprinting and crosslinking.

Our approach involves system integration of instrumentation, organic chemistry, molecular biology, electrophoresis and software to the task of increasing sequencing accuracy and efficiency. Likewise we integrate such instruments and others with the needs of acquiring and annotation of large-scale microbial and human genomic sequence and population polymorphisms.

To establish functions for new genes, we use large scale phenotyping by multiplexed growth competition assays, both by targeted deletion and by saturation insertional mutagenesis. We will continue to develop a system to sequence DNA using electrophore mass-tags (EMTs). We will establish genome-scale experimental methods for sequence annotation.

The most significant findings in 1995-1996 were 1) Demonstration of use of electrophore mass-tags in dideoxy sequencing. 2) Development of IR-laser desorption method and model. 3) A novel dsDNA microarray synthesis strategy. 4) A new amplifiable differential display for whole-genome in vivo DNA-protein interactions. 5) Establishment and application of a microbial DNA-protein interaction database.

DOE Grant No. DE-FG02-87ER60565.

A PAC/BAC End-Sequence Data Resource for Sequencing the Human Genome: A 2-Year Pilot Study

Pieter de Jong
Roswell Park Cancer Institute; Buffalo, NY 14263
716/845-3168, Fax: -8849, *pieter@dejong.med.buffalo.edu*
http://bacpac.med.buffalo.edu

Large scale sequencing of the Human genome requires the availability of high-fidelity clones with large genomic inserts and a mechanism to find clones with minimal overlaps within the clone collections. The first need can be satisfied with bacterial artificial chromosome libraries (PACs and BACs) which already exist and further such libraries now being developed. However, a cost-effective way for establishing high-resolution contig maps for the human genome has not yet been established. Recently, a new approach for virtual screening for overlapping clones has been proposed by several research groups and has been discussed eloquently in a manuscript by Venter et al., 1996 (Nature). We will implement this approach for use with our human PAC and BAC libraries and use the first year as a pilot stage. The goal of the one year pilot is to prove the feasibility of large scale end sequencing and to demonstrate usefulness.

The first goal will be met by sequencing the ends for 40,000 clones from our existing PAC library and from BAC libraries currently being developed under NIH funding within our laboratory. The end-sequencing will be based on our new DOP-vector PCR procedure (Chen et al, 1996, Nucleic Acids Research 24, 2614-2616). All sequence data will be made available through public databases (GSDB, GDB, Genbank) and will also become BLAST searchable through the UTSW WWW site from our collaborator, Glen Evans. In view of our current under-developed informatics structure, we do not expect to provide BLAST search access through our own web site during the pilot phase.

To prove the usefulness of available end sequences, we will prepare a chromosome 14-enriched clone collection from our current 20-fold deep PAC library. To detect the chromosome 14 clones, we will use as hybridization probes a set of 1,000 mapped STS markers available from Paul Dear (MRC, Cambridge, UK), the about 600 markers present in the Whitehead map and the in situ mapped BAC and PAC clones available from Julie Korenberg. We will hybridize with these existing markers in probe pools, specific for regions of chromosome 14. Thus we will isolate region-enriched PAC clone collections.

Assuming that the clone collections will be at least 50%-specific for chromosome 14 (50% false positives) and will include most of the chromosome 14 PACs from our library, a collection of about 35,000 clones is expected.

Hence, the bulk of the end sequences obtained during the first year will be derived from the chromosome 14 enriched set and should result in a sequence ready clone collection covering about 100 Mbp of the human genome. The purity of the chromosome 14 PAC collection will be characterized in a number of different ways, including testing with independent markers not used as probes and by FISH analysis of a representative set of PAC clones. To test the usefulness of the end sequence resource, the Sanger Centre will sequence chromosome 14 PACs from our collection and identify overlapping clones by virtual screening, using our end-sequence database.

If overlapping clones can not be found with the expected level of redundancy in the end-sequence database, we will screen the original PAC library with probes or STS markers derived from the sequenced PAC clones.

Subcontract under Glen Evans' DOE Grant No. DE-FC03-96ER62294.

Multiple-Column Capillary Gel Electrophoresis

Norman Dovichi
Department of Chemistry; University of Alberta;
Edmonton, Alberta, Canada T6G 2G2
403/492-2845, Fax: -8231, *norm.dovichi@ualberta.ca*
http://hobbes.chem.ualberta.ca

The objective of this project is to develop high-throughput DNA sequencing instrumentation. A two-dimensional arrayed capillary electrophoresis instrument is under development.

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.

DOE Grant No. DE-FG02-91ER61123.

DNA Sequencing with Primer Libraries

John J. Dunn, Laura-Li Butler-Loffredo, and **F. William Studier**
Biology Department; Brookhaven National Laboratory;
Upton, NY 11973
516/344-3012, Fax: -3407, *dunn@genome1.bio.bnl.gov*
http://genome5.bio.bnl.gov

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 M 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.

DOE Grant No. DE-AC02-76CH00016.

References

(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).

Rapid Preparation of DNA for Automated Sequencing

John J. Dunn, Matthew Randesi, and **F. William Studier**
Biology Department; Brookhaven National Laboratory; Upton, NY 11973
516/344-3012, Fax: -3407, *dunn@genome1.bio.bnl.gov*
http://genome5.bio.bnl.gov

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 lambda 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 lambda (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.

DOE Grant No. DE-AC02-76CH00016.

A PAC/BAC End-Sequence Database for Human Genomic Sequencing

Glen A. Evans, Dave Burbee, Chris Davies, Trey Fondon, Tammy Oliver, Terry Franklin, Lisa Hahner, Shane Probst, and Harold R. (Skip) Garner
Genome Science and Technology Center and McDermott Center for Human Growth and Development; University of Texas Southwestern Medical Center at Dallas; Dallas, TX 75235-8591
214/648-1660, Fax: -1666, *gevans@swmed.edu*
http://mcdermott.swmed.edu

While current plans call for completing the human genome sequence in 2003, major obstacles remain in achieving the speed and efficiency necessary to complete the task of mapping and sequencing. As an approach to this problem, we proposed a novel approach to large scale construction of sequence-ready physical clone maps of the human genome utilizing end-specific sequence sampling. An earlier pilot project was initially carried out to develop a GSS (genomic sequence sampled) map of human chromosome 11 by sequencing the ends of 17,952 chromosome 11 specific cosmids. This chromosome 11-specific end-sequence database allows rapid and sensitive detection of clone overlaps for chromosome 11-sequencing.

In this project, we propose to evaluate the utility of PAC and BAC end-sequences representing the entire human genome as a tool for complete, high accuracy mapping and sequencing. In this approach, we utilized total genomic PAC/BAC libraries (constructed by P. de Jong, RPCI), followed by end-sequencing of both ends of each clone in the library and limited regional mapping of a subset of clones as sequencing nucleation points by FISH (Fluorescence in situ hybridization).

To initiate regional analysis, a single clone would be sequenced by shotgun or primer directed sequencing, the entire sequence used to search the end-database for overlapping clones, and the minimal overlapping clones for extending the sequence selected. This approach would allow rational and efficient simultaneous mapping and sequencing, as well as expediting the coordination and exchange of information between large and small groups participating in the human genome project.

In this pilot project proposal we are carrying out automated end-sequencing of approximately 40,000 PAC and BAC clones representing the entire human genome, as well as about 500 PAC clones localized to human chromosomes 11 and 15. The clones and resulting end-sequence data base will be utilized to 1) nucleate regions of interest for large scale sequencing concentrating on regions of chromosome 11 and 15, 2) correspond with regions mapped by other methods to confirm the mapping accuracy and 3) used to evaluate the use of random clone end sequence libraries. DNA sequencing is being carried out in an entirely automated fashion using a Beckman/Sagian robotic system, ABI 377 automated sequencers and automated sequence data processing, annotation and publication using a Hewlett Packard/Convex superparallel computer located at the UTSW genome center. FISH analysis of a sample of PAC clones has been carried out and defines the potential chimera rate in existing PAC libraries as less than 1.2%. This effort will be coordinated with efforts of other groups carrying out PAC and BAC library construction, PAC and BAC end-sequencing and FISH analysis to avoid duplication of effort and provide a comprehensive end-sequence library and data set for use by the international human genome sequencing effort.

DOE Grant No. DE-FC03-96ER62294.

Automated DNA Sequencing by Parallel Primer Walking

Glen A. Evans, Dave Burbee, Chris Davies, Jeff Schageman, Shane Probst, Terry Franklin, Ken Kupfer, and Harold R. (Skip) Garner
Genome Science and Technology Center and McDermott Center for Human Growth and Development; University of Texas Southwestern Medical Center at Dallas; Dallas, TX 75235-8591
214/648-1660, Fax: -1666, *gevans@swmed.edu*
http://mcdermott.swmed.edu

The development of efficient mapping approaches coupled with high throughput, automated DNA sequencing remains one of the key challenges of the Human Genome Project. Over the past few years, a number of strategies to expedite clone-by-clone DNA sequencing have been developed including efficient shotgun sequencing, sequencing of nested deletions, and transposon-mediated primer insertion. We have developed a novel sequencing strategy applicable to high throughput, large scale genomic analysis based upon DNA sequencing directly primed on of cosmid templates using custom-designed, automatically synthesized oligonucleotide primers. This approach of directed primer "walking" would allow the number of sequencing reactions and the efficiency of sequencing to be vastly improved over traditional shotgun sequencing.

Custom primer design has been carried out using software we developed for prediction of "walking" primers directly from the output of ABI377 automated DNA sequencers, and the output used to automatically program synthesis of the custom primers using 96 or 192 channel oligonucleotide synthesizers constructed at UTSW. Automated operation of the sequencing system is thus possible where results of each sequencing reaction is used to predict, synthesize, and carry out appropriate extension reactions for downstream "walking". A automated prototype system has been assembled where dye terminator DNA sequencing can be carried out from 96 cosmid templates simultaneously followed by prediction of oligonucleotide "walking" primers for extending the sequence of each fragment, and programming an attached 96-channel oligonucleotide synthesizer to initiate a second round of sequencing. Using a set of nested cosmids covering 800 kb at 5X redundancy, primer directed sequencing should allow completion of 800 kb of finished, high accuracy DNA sequence in 8 to 16 cycles. Furthermore, coupling of automated DNA sequencing instrumentation to DNA sequence analysis programs and multichannel oligonucleotide synthesizers will allow almost complete automation of sequencing process and the development of instrumentation for completely unattended DNA sequencing.

DOE Grant No. DE-FG03-95ER62055.

*Parallel Triplex Formation as Possible Approach for Suppression of DNA-Viruses Reproduction

V.L. Florentiev, A.K. Shchyolkina, I.A. Il'icheva, E.N. Timofeev, and S. Yu Tsybenko
Engelhardt Institute of Molecular Biology; Russian Academy of Sciences; Moscow 117984, Russia
Fax: +7-095/135-1405, *flor@imb.imb.ac.ru*

It is well known that homopurine or homopyrimidine single stranded oligonucleotides can bind to homopurine-homopyrimidine sequences of two-stranded DNA to form stable three-stranded helices. In such triplexes two identical strands have antiparallel orientation. We denote these triplexes as "antiparallel" or "classical" triplexes.

A particular interest of investigators to triplexes has arisen due to an elegant idea of using triplexes as sequence-specific tools for purposeful influence on DNA duplexes. Triplex forming oligonucleotides were shown to be potentially useful as regulators of gene expression and subsequently as therapeutical (antiviral) agents.

A significant limitation to the practical application of antiparallel triplex is the requirement for homopurine tracts in target DNA sequences. Numerous investigations slightly

expanded the repertoire of triple-forming sequences but did not completely remove this limitation.

It was recently shown that during homologous recombination promoted by RecA a triple-stranded DNA intermediate was formed. Such a structure is a new form of the triple helix. In sharp contrast with the "classical" triplexes their third strand is parallel to the identical strand of the Watson-Crick duplex. We denote this structure as "parallel" triplex. Recently, the parallel triplex was obtained only by deproteinization of joint molecules generated by recombination proteins.

We first obtained experimental (chemical probe, melting curves and fluorescence due binding) results that provide convincingly evidence for protein-independent formation of parallel triplex [1] and than confirmed this fact by FTIR data [2]. Because the parallel triplex can be formed for any sequence, it might be "ideal" potential tool for sequence specific recognition of DNA. Unfortunately, low stability of parallel triplexes prohibits practical application of these structures.

Earlier we found that propidium iodide stabilizes selectively the parallel triplexes [3]. This fact was the basis of new approach to stabilization of parallel triplexes being developed by us now. The approach consists in use of targeting oligonucleotide, which contains in internucleotide linkage the alkyl insert coupled with intercalated ligand through linker. Length of linker was chosen to allow ligand to intercalate in the same stacking-contact (length of linker was picked by molecular dynamic calculations).

Preliminary study showed that presence of intercalating inserts increase considerably stability of DNA duplexes [4]. Now we are investigating in detail effect of such modification of targeting oligonucleotides on stability of parallel triplexes.

DOE Grant No. OR00033-93CIS005.

References

1. Shchyolkina, A. K., Timofeev, E. N., Borisova, O. F., Il'icheva, I.A., Minyat, E. E., Khomyakova, E. B. and Florentiev, V. L. (1994) The R-form DNA does exist. FEBS Letters, 339, 113-118.
2. Dagneaux, C., Gousset, H., Shchyolkina, A. K., Ouali, M., Lettelier, R., Liquier, J., Florentiev, V. L. and Taillander, E. (1996) Parallel and antiparallel AA-T intramolecular triple helices. Nucleic Acids Res., 24, 4506-4512.
3. Borisova, O. F., Shchyolkina, A. K., Timofeev, E. N., Tsybenko, S. Yu., Mirzabekov, A. and Florentiev, V. L. (1995) Stabilization of parallel triplex with propidium iodide. J. Biomol. Struct. Dynam., 13, 15-27.
4. Timofeev, E. N., Smirnov I. P., Haff, L. A., Tishchenko, E. I., Mirzabekov, A. D. and Florentiev, V. L. (1996) Methidium intercalator inserted into synthetic oligonucleotides. Tetrahedron Lett., 37, 8467-8470.

Advanced Automated Sequencing Technology: Fluorescent Detection for Multiplex DNA Sequencing

Andy Marks, Tony Schurtz, F. Mark Ferguson, Leonard Di Sera, Alvin Kimball, Diane Dunn, Doug Adamson, Peter Cartwright, Robert B. Weiss,[1] and **Raymond F. Gesteland**[1]
Department of Human Genetics and [1]Howard Hughes Medical Institute; University of Utah; Salt Lake City, UT 84112
Gesteland: 801/581-5190, Fax: /585-3910
ray.gesteland@genetics.utah.edu

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 sys-

tem, 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.

DOE Grant No. DE-FG03-94ER61817.

Reference

[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.

Resource for Molecular Cytogenetics

Donna Albertson, Colin Collins, **Joe Gray**,[1] Steven Lockett, **Daniel Pinkel**,[1] Damir Sudar, Heinz-Ulrich Weier, and Manfred Zorn
Lawrence Berkeley National Laboratory; Berkeley, CA 94720 and [1]University of California; San Francisco, CA 94143
Gray: 415/476-3461, Fax: -8218, *gray@cc.ucsf.edu*
Pinkel: 415/476-3659, Fax: -8218, *pinkel@cc.uscf.edu*
http://rmc-www.lbl.gov

The purpose of the Resource for Molecular Cytogenetics is to develop molecular cytogenetic techniques, instruments and reagents needed to facilitate large scale genomic DNA sequencing and to assist in identification and functional characterization of genes involved in disease susceptibility, genesis and progression. This work is closely coordinated with the LBNL Human Genome Program and directly supports research in the LBNL Life Sciences Division and the UCSF Cancer Center. Work currently is in four areas: a)Genome analysis technology, b)Probe development and physical map assembly, c)Digital imaging microscopy and d)Informatics. The Resource acts as a catalyst for research in several areas so some support comes from Industry, the NIH and NIST.

Probe development and physical map assembly: The Resource maintains a list of over a thousand publicly available probes suitable for molecular cytogenetic studies. These include approximately 600 probes each selected by the Resource to contain a known STS or EST. Probes selected by the Resource can be requested through our web page.

The Resource also participates in the development of low and high resolution physical maps to facilitate analysis and characterization of genetic abnormalities associated with human disease. Low resolution mapping panels with probes distributed at few megabase intervals have been completed this year for chromosomes 1, 2, 3, 7, 8, 10, and 20. The mapped STSs associated with these probes facilitate movement from low to high resolution physical maps. STS content mapping and DNA fingerprinting have been applied to develop a high resolution, sequence-ready map comprised of BAC and P1 clones for the ~1Mb region of chromosome 20 between WI9227 and D20S902. This region is amplified in ~10% of human breast cancers. Approximately 300 kb of this region has been sequenced by the LBNL Human Genome Program.

Quantitative DNA fiber mapping (QDFM) has been developed this year to facilitate high resolution analysis of genomic overlap between cloned probes. In this approach, cloned DNA molecules are uniformly stretched during drying by the hydrodynamic action of a receding meniscus. The position of specific sequences along the stretched DNA molecules is visualized by fluorescence in situ hybridization (FISH) and measured by digital image analysis. QDFM has been used to map gamma alpha transposons, plasmid or cosmid probes along P1 molecules, and P1 or PAC clones along straightened YAC molecules with few kilobase resolution. QDFM is now being studied to determine its utility in the assembly of minimally overlapping, sequence-ready contigs, assessment of the integrity of cloned BACs and mapping of subclones prepared for directed DNA sequencing along the clone from which they were derived.

Genome analysis technology: The Resource has participated in the development of comparative genomic hybridization (CGH) as a tool for detection and mapping of changes in relative DNA sequence copy number in humans and mouse. This year, CGH to arrays of cloned probes (CGHa) has been demonstrated. This is advantageous because it allow aberrations to be mapped with resolution determined by the genomic spacing of probes on the array. CGHa also is attractive since it appears to be linear over a relative copy number range of at least 104 between the two nucleic acid samples being compared.

The Resource has participated in the development of FISH approaches to analysis of relative gene expression in normal and aberrant tissues. FISH with cloned or predicted expressed sequences, previously developed in C. elegans, is now being applied to the assessment of expression of human genes. The C. elegans work suggests a throughput of several dozen sequences per month. Information from this approach will be important in assessment of the function of newly discovered genes, including those predicted from DNA sequencing.

(abstract continued)

Digital imaging microscopy: The Resource supports work in microscopy, image processing and analysis methods needed for CGH and CGHa, 3D FISH, tissue analysis, rare event detection, multi-color image acquisition, aberration scoring for biodosimetry, and analysis of FISH to DNA fibers. Developments this year include an improved package for CGH and prototype systems for analysis of DNA fibers, CGHa arrays and semiautomatic segmentation of nuclei in three dimensions.

Informatics: The Resource maintains a web site at http://rmc-www.lbl.gov that summarizes information about mapped probes. Probes developed by the Resource can be requested directly through this page. In addition, the Resource has developed a Web page for exchange of genomic, genetic and biologic information between geographically disperse collaborators. The page, under password control, carries information about physical maps, genomic sequence, sequence annotation, and gene expression images.

DOE Contract No. DEAC0376SF00098.

DNA Sample Manipulation and Automation

Trevor Hawkins
Center for Genome Research; Whitehead Institute/Massachusetts Institute of Technology; Cambridge, MA 02139
617/252-1910, Fax: -1902, *tlh@genome.wi.mit.edu*
http://www-genome.wi.mit.edu

The objective of this project is to develop a high-throughput, fully automated robotic device for the complete automation of the sequencing process. We also aim to further develop DNA sequencing electrophoresis systems and to integrate these devices with our robotics.

We have built the Sequatron, an integrated, robotic device which automates the tasks of DNA purification and setup of thermal cycle sequencing reactions. 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 or 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 throughout 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 1g 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.

DOE Grant No. DE-FG02-95ER62099.

Relevant Publication

DeAngelis, M., Wang, D., & Hawkins, T. (1995) Nucl. Acids. Res 23, 4742-4743.

Construction of a Genome-Wide Characterized Clone Resource for Genome Sequencing

Leroy Hood, **Mark D. Adams**,[1] and **Melvin Simon**[2]
University of Washington; Seattle, WA 98195-7730
206/616-5014, Fax: /685-7301, *tawny@u.washington.edu*
[1]The Institute for Genomic Research; Rockville, MD 20850; *mdadams@tigr.org*
[2]California Institute of Technology; Pasadena, CA 91125; *simonm@starbase1.caltech.edu*

Bacterial artificial chromosomes (BACs) represent the state of the art cloning system for human DNA because of their stability and ease of manipulation. Venter, Smith and Hood (Nature 381:364-366, 1996) have proposed a strategy based on the use of sequences from the ends of all clones in a deep coverage BAC library to produce a sequence-ready set of clones for the human genome. We propose to demonstrate the effectiveness of this strategy by performing a directed test, initially on chromosomes 16 and 22, and continuing on to chromosome 1. All available markers on chromosome 16 (including the large number of soon-to-be-available radiation hybrid markers) will be used to screen the existing 8x BAC library at CalTech. This will serve to evaluate the quality of the library in terms of representation of broad chromosomal regions. A similar procedure will be used for chromosome 22, except that the existing BAC map will be used to select more evenly spaced markers for screening, including use of end-sequence markers from the current chromosome 22 BAC map constructed in the Simon lab. Each identified clone will be rearrayed from the library and end sequenced. This information will dovetail nicely with ongoing sequencing projects at TIGR and the Sanger Centre, which will in turn provide additional information on the average degree of BAC overlap detectable by this method, the degree of interference with genome-wide repeats, and the appropriate use of fingerprinting as an early or late addition to the end-sequencing information. In addition, we will develop and implement cost-effective, high-throughput methods of preparing and end-sequencing BAC DNA that are suitable for scaling to characterization

of the full 400,000 clones necessary for characterization of a 15x human BAC library.

DOE Grant No. DE-FC03-96ER62299.

DNA Sequencing Using Capillary Electrophoresis

Barry L. Karger
Barnett Institute; Northeastern University; Boston, MA 02115
617/373-2867 or -2868, Fax: -2855
bakarger@lynx.neu.edu

During the past year, we have made major progress in the design of a replaceable polymer matrix for DNA sequencing and the development of the first generation multiple capillary array of 12 capillaries. We also implemented ultrafast separation of dsDNA (e.g. 30 sec for complete resolution of the standard X174-HAE III restriction fragments).

In the separation of sequencing reaction products, we completed a study on the role of polymer molecular weight and concentration. Using linear polyacrylamide (LPA), the polymer with which we have had our most success, we have achieved 1000 base read lengths in 1 1/2 hrs. Optimization of column length, electric field and column temperature (50° C) was required. Using emulsion polymerization, we are now able to produce LPA powders with MW of ~10^4 k Da. The fully replaceable matrix is very powerful for rapid sequencing of long reads.

We have successfully implemented a 12-capillary array instrument and are using it to study issues of ruggedness in routine sequencing. As part of this, we have developed a sample clean-up procedure which reduces all reactions to a similar state in terms of sample solution prior to injection. The results of this work have led to the design of a 96-capillary array that we will implement over the next year.

We have also achieved very fast separations of ss- and dsDNA using short capillaries and very high yields. For example, sequencing 300 bases in 3–4 mins. has been shown, as well as very rapid mutational analysis. Implementation of such speeds on a capillary array will create an instrument for high throughput automated analysis.

DOE Grant No. DE-FG02-90ER60985.

Ultrasensitive Fluorescence Detection of DNA

Richard A. Mathies and **Alexander N. Glazer**
Departments of Chemistry and Molecular and Cell Biology; University of California; Berkeley CA 94720
510/642-4192, Fax: -3599, *rich@zinc.cchem.berkeley.edu*

The overall goal of this project is to develop new fluorescence labeling methods, separation methods and detection technologies for DNA sequencing and genomic analysis.

Highlights along with representative publications are given below.

Energy Transfer Primers. Families of sequencing and PCR primers have been developed that contain both fluorescence donor and acceptor chromophores.[1] These labeled primers with optimized excitation and emission properties provide from 2- to 20-fold enhanced signal intensities in automated DNA sequencing with slab gels and with capillary arrays.[2] The reduced spectral cross talk of these ET primers also makes them valuable in PCR product and STR analyses.[3]

New Intercalation Dye Labels. A new family of heterodimeric bis-intercalation dyes has been synthesized exploiting the concept of fluorescence energy transfer between two different cyanine intercalators.[4] By tailoring the spectroscopic properties of the dyes, labels with intense emission above 650 nm following 488 nm excitation have been fabricated. By adjusting the spacing linker between the two dyes, the binding affinity has also been optimized. These molecules are useful for noncovalent multiplex labeling of ds-DNA in a wide variety of multicolor analyses.[5]

Capillary Electrophoresis Chips. Capillary and capillary array electrophoresis systems have been photolithographically fabricated on 2x3' glass substrates.[6] These devices provide high quality electrophoretic separations of ds-DNA fragments and DNA sequencing reactions with a 10-fold increase in speed.[7] Arrays of up to 32 capillaries on a single chip have been fabricated.

Single DNA Molecule Fluorescence Burst Detection. A confocal fluorescence system has been used to demonstrate that single molecule fluorescence burst counting can be used to detect CE separations of ds-DNA fragments. Fragments as small as 50 bp can be counted and mass sensitivities as low as 100 molecules per electrophoresis band are possible. This technology should be valuable in incipient cancer and trace pathogen detection.[8]

DOE Grant No. DE-FG03-91ER61125.

(abstract continued)

References

1. 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).
2. Ju, J., Glazer, A. N. and Mathies, R. A. Energy Transfer Primers: A New Fluorescence Labeling Paradigm for DNA Sequencing and Analysis, Nature Medicine 2, 180-182 (1996).
3. 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).
4. Benson, S. C., Zeng, Z., and Glazer, A. N. Fluorescence Energy Transfer Cyanine Heterodimers with High Affinity for Double-Stranded DNA. I. Synthesis and Spectroscopic Properties, Anal. Biochem. 231, 247-255 (1995).
5. Zeng, Z., Benson, S. C., and Glazer, A. N. Fluorescence Energy Transfer Cyanine Heterodimers with High Affinity for Double-Stranded DNA. II. Applications to Multiplex Restriction Fragment Sizing, Anal. Biochem. 231, 256-260 (1995).
6. 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).
7. Woolley, A. T. and Mathies, R. A. Ultra-High-Speed DNA Sequencing Using Capillary Array Electrophoresis Chips, Analytical Chemistry 67, 3676-3680 (1995).
8. 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).

Joint Human Genome Program Between Argonne National Laboratory and the Engelhardt Institute of Molecular Biology

Andrei Mirzabekov,[1,2] G. Yershov,[1,2] Y. Lysov,[2] V. Barsky,[2] V. Shick,[2] and S. Bavikin[1]
[1]Argonne National Laboratory; Argonne, Il 60439
630/252-3161 or -3361, Fax: /252-3387
amir@everest.bim.anl.gov
[2]Engelhardt Institute of Molecular Biology; 117984 Moscow, Russia

In 1996, more than thirty U.S. and Russian research workers participated in the joint Human Genome Program between Argonne National Laboratory and Engelhardt Institute of Molecular Biology on the development of sequencing by hybridization with oligonucleotide microchips (SHOM).

During this year, about twenty Russian scientists have been working from 3 months to 1 year in ANL. In this period, 3 papers have been published and 5 papers accepted for publication, 3 more papers are submitted for publication.

The main research efforts of the group have been concentrated in three directions:

I. Improvement of SHOM technology.
II. Development of SHOM for the needs of Human Genome Program.
III. Development of new approaches based on SHOM technology.

I. Improvement of SHOM technology

As a major result of the work in this direction, simple, reliable and effective methods of microchip manufacturing, sample preparations, and quantitative hybridization analysis by fluorescence microscopy have been developed or improved.

1. Photopolymerization technique for production of micromatrices of polyacrylamide gel pads on hydrophobicized glass surface was improved to become a simple, highly reproducible and inexpensive procedure (7).

2. New and cheaper chemistry of the oligonucleotide immobilization has been developed and introduced for production of more durable microchips. It is based on the use of amino-oligonucleotides and aldehyde-gels instead of 3-methyluridine-oligonucleotides and hydrazide-gels (3).

3. Four-pin robot has been constructed with computer control of every microchip element production. High quality microchips with 4100 immobilized oligonucleotides have been manufactured and the complexity of the microchips can easily be scaled up to a few tens of thousand elements.

4. Two-color fluorescence microscope has been equipped for regular use with proper mechanics and software. It allows investigators to regularly use the automatic quantitative monitoring of the hybridization on the whole microchip and to measure the kinetics of hybridization as well as the melting curves of duplexes formed with all microchip oligonucleotides (1,2,8).

5. Four-color fluorescence microscope was manufactured and four proper fluorescence dyes are at present under selection.

6. Chemical methods of introduction of several fluorescence dyes into DNA and RNA with or without fragmentation have been developed and regularly used in SHOM experiments (4).

7. A theory describing the kinetics of hybridization with gel-immobilized oligonucleotides has been developed (5).

8. Simple and relatively inexpensive equipment (around $10,000 per set) has been produced for manual manufacturing of microchips and fluorescence measurement of hybridization, which will enable every laboratory to produce and practically use microchips containing up to 100 immobilized oligonucleotides or other compounds.

II. Application of SHOM

Although the main goal of our SHOM development is to produce a simple de novo sequencing procedure, a number

of other SHOM applications have been tested as intermediate steps in the SHOM research.

1. Sequence analysis and sequencing

A number of technical problems should be solved for de novo sequencing although they are much less stringent for comparative sequence analysis than for de novo sequencing. Among these:

a) Reliable discrimination of perfect and mismatched duplexes. We have significantly improved the discrimination by decreasing the length of hybridized oligonucleotides to 6-and 8-mers (1, 7) and by using 5-mers in "contiguous stacking" hybridization (1,2). Essential improvement was also achieved by automatic measuring of the melting curves for duplexes formed in each microchip element and calculating their thermodynamic parameters, free energy, enthalpy and entrophy for different regions of the melting curves and by comparing them with these parameters for perfect duplexes. In addition, a highly reliable discrimination was achieved by using two-color fluorescence microscopy and by quantitative comparison of the hybridization pattern of a known DNA or synthetic oligonucleotides and DNA under study labeled with different fluorophores (8).

b) Difference in hybridization efficiency depends on the GC-content and the length of the duplex. We have equalized the efficiency by choosing proper concentration for the immobilized oligonucleotide (6,7) and also by increasing the effective length of immobilized oligonucleotides by adding at one or both their ends 5-nitroindole as a universal base or a mixture of four bases (2).

c) Interference of hairpins and other structures in DNA with less stable duplexes formed upon the DNA hybridization with comparatively short immobilized oligonucleotides of the microchip. This interference was decreased by fragmentation of the analysed sample of DNA and RNA in the course of incorporation of a fluorescence label (4). We have also tested incorporation by a chemical bond of an intercalator into immobilized oligonucleotides that stabilized its base paring with DNA over hairpin formation (10).

d) Necessity to increase the microchip complexity for sequencing long DNA stretches. As an alternative, further development of so-called contiguous stacking hybridization was shown to improve the efficiency of 8-mer microchip up to that of 13-mer microchip so that DNA of several kilobases in length could be sequenced by SHOM (2).

e) 6-mer microchips for sequencing and sequence analysis. We have now come to the stage of manufacturing microchips containing 4,096 (i.e. all possible) 6-mers. The control tests partly described above have shown that these microchips can be effectively used for sequence analysis, mutation diagnostics and detection of sequencing mistakes by conventional gel-sequencing methods. We hope that after demonstrating the efficiency of 6-mer microchips, we shall be able to get sufficient financial support for production of the microchip with all 65,536 8-mers.

2. Mutation diagnostics and gene polymorphism analysis

The improvements described above have been introduced for reliable ("Yes" or "No" mode) identification of single-base changes in human genomic DNA. The efficiency of SHOM has been demonstrated for identification of a number of b-thalassemia mutations (1,2,8) and HLA allele variations in the human genome.

3. Identification of microorganisms and gene expression monitoring

Bacterial microchips have been manufactured and tested. Their ability for reliable identification of a number of bacterial strains in the sample has been demonstrated (6). The chips containing oligonucleotides complementary to specific regions of 16S ribosomal RNA were hybridized with samples of rRNA, total RNA, DNA and RNA transcripts of PCR-amplified genomic rDNA. Similar preliminary experiments demonstrated the efficiency of SHOM for monitoring the gene expression.

III. Development of new approaches based on the SHOM technology

1. Enzymatic modification of nucleic acids on selected elements of the oligonucleotide chip. The gel pads of the oligonucleotide chip are separated from each other by hydrophobic glass surface. It prevents the cross-talking of the chip elements when a drop of solution is applied on specified elements. At the same time, a high porosity of the gel allows diffusion of large proteins into the gel. We have demonstrated that immobilized oligonucleotides can be enzymatically phosphorylated and ligated with contiguously stacked 5-mer after hybridization with DNA. A walking sequencing procedure by stacked pentanucleotides was proposed that is based on enzymatic ligation and phosphorylation on oligonucleotides chips (9).

2. DNA fractionation on oligonucleotide chips. Due to the same properties, the oligonucleotide chips are used for fractionation of DNA after DNA hybridization with some complementary oligonucleotides of the chip. A new procedure for sequencing long DNA pieces was proposed that is based on fractionation of DNA on fractionating oligonucleotide chips followed by sequencing of the isolated DNA by SHOM on sequencing microchips. The procedure allows the investigator to skip cloning and mapping of long DNA pieces (9).

Conclusions

It appears that the major technical problems of SHOM have been in most part solved, and this technology can al-

ready be applied for sequence analysis and checking the accuracy of conventional sequencing methods. A number of other applications in the Human Genome Program are within the reach of SHOM, such as mutation screening, gene polymorphism studies, detection of microorganisms, gene expression studies, etc. Application of SHOM for de novo DNA sequencing requires manufacturing of more complicated microchips and improvement of some other, already available methods.

DOE Contract No. W-31-109-Eng-38.

References

1. Yershov G., Barsky V., Belgovsky A., Kirillov Eu., Kreindlin E., Ivanov I., Parinov S., Guschin D., Drobishev A., Dubiley S., Mirzabekov A. DNA analysis and diagnostics on oligonucleotide microchips. // Proc. Natl. Acad. Sci. 1996. Vol. 93. 4913-4918.
2. Parinov S., Barsky V., Yershov G., Kirillov Eu., Timofeev E., Belgovskiy A., Mirzabekov A. DNA sequencing by hybridization to microchip octa-and decanucleotides extended by stacked pentanucleotides. // Nucl. Acids Res. 1996. Vol. 24. N 15. P. 2998-3004.
3. Timofeev E., Kochetkova S., A., Mirzabekov A. Radioselective immobilization of short oligonucleotides to acrylic copolymer gels // Nucl. Acids Res. 1996. Vol. 24. N 16. P. 3142-3148.
4. Prudnikov D., Mirzabekov A. Chemical methods of DNA and RNA fluorescent labelling. // Nucl. Acids Res. 1996., in press.
5. Livshits M., Mirzabekov A. Theoretical analysis of the kinetics of DNA hybridization with gel-immobilized oligonucleotides. // Biophys. J. 1996. Vol. 71, in print//.
6. Guschin D., Mobarry B., Proudnikov D., Stahl D., Rittmann B., Mirzabekov A. Oligonucleotide microchips as genosensors for determinative and environmental studies in microbiology //Applied and Environmental Microbiology, in print//.
7. Guschin D., Yershov G., Zaslavsky A., Gemmell A., Shick V., Lysov Yu., Mirzabekov A. A simple method of oligonucleotide microchip manufacturing and properties of the microchips // submitted for publication.
8. Drobyshev A., Mologina N., Shik V., Pobedimskaya D., Yershov G., Mirzabekov A. Sequence analysis by hybridization with oligonucleotide microchip: identification of beta-thalassemia mutations // Gene (in print).
9. Dubiley S., Kirillov Eu., Lysov Yu., Mirzabekov A. DNA fractionation, sequence analysis and ligation of immobilized oligomers on oligonucleotide chips // submitted for publication.
10. Timofeev E., Smirnov I.P., Haff L.A., Tishchenko E.I., Mirzabekov A.D., Florentiev V.L.. Methidium Intercalator Inserted into Synthetic Oligonucleotides // Tetrahedron Letters 1996, v.37, N47, p.8467.

Relevant Publication

Methods of DNA sequencing by hybridization based on optimizing concentration of matrix-bound oligonucleotide and device for carrying out same by Khrapko K., Khorlin A., Ivanov I., Ershov G., Lysov Yu., Florentiev V., Mirzabekov A. US Patent 5,552,270, Sep. 3, 1996. PCT/RU92/00052, filed Mar 18, 1992.

High-Throughput DNA Sequencing: SAmple SEquencing (SASE) Analysis as a Framework for Identifying Genes and Complete Large-Scale Genomic Sequencing

Robert K. Moyzis and Jeffrey K. Griffith[1]
Center for Human Genome Studies; Los Alamos National Laboratory; Los Alamos, NM 87545
505/667-3912, Fax: -2891, *moyzis@telomere.lanl.gov*
[1]University of New Mexico; Albuquerque, NM 87131

The human chromosome 5 and 16 physical maps (Doggett et al., Nature 377:Suppl:335-365, 1995; Grady et al., Genomics 32:91-96, 1996) provide the ideal framework for initiating large-scale DNA sequencing. These physical mapping studies have shown clearly that gene density in humans will vary greatly. For example, band 16q21, consisting of 8 Mb of DNA, has no genes or trapped exons assigned to it, as yet. In contrast, band 16p13.3 has an extremely high density of coding regions in the DNA examined to date (i.e., multiple genes/cosmid). Given this wide variation in gene density and current sequencing costs, we propose that newly targeted genomic regions should be analyzed first by a "Lewis and Clark" exploratory approach, before committing to full length DNA sequencing. We are using a SAmple SEquencing (SASE) approach to rapidly generate aligned sequences along the chromosome 5 and 16 physical maps. 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 (for example, primer walking or transposon sequencing). SASE analysis has been initiated on the 40 Mb short arm of chromosome 16 and the 45 Mb short arm of chromosome 5. We propose to make 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, will allow for the effective "democratization" of the genome, allowing numerous laboratories to share and contribute to the growing genome databases.

DOE Grant No. DE-FG03-96ER62298.

One-Step PCR Sequencing

Kenneth W. Porter, J. David Briley, and **Barbara Ramsay Shaw**
Department of Chemistry; Duke University; Durham, NC 27708
919/660-1553, Fax: -1605, *ken@chem.duke.edu*

A method is described to simultaneously amplify and sequence DNA using a new class of nucleotides containing boron. During the polymerase chain reaction, boron-modified nucleotides, i.e. 2'-deoxynucleoside 5'-a-[P-borano]-triphosphates,[1,2] are incorporated into the product DNA. The boranophosphate linkages are resistant to nucleases and thus the positions of the boranophosphates can be revealed by exonuclease digestion, thereby generating a set of fragments that defines the DNA sequence. The boranophosphate method offers an alternative to current PCR sequencing methods.

Single-sided primer extension with dideoxynucleotide chain terminators is avoided with the consequence that the sequencing fragments are derived directly from the original PCR products. Boranophosphate sequencing is demonstrated with the Pharmacia and the Applied Biosystems 373A automatic sequencers producing data that is comparable to cycle sequencing.

DOE Grant No. DE-FG02-97ER62376 and NIH Grant No. HG00782.

References

[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.

Automation of the Front End of DNA Sequencing

Lloyd M. Smith and **Richard A. Guilfoyle**
University of Wisconsin; Madison, WI 53706
Guilfoyle: 608/265-6138, Fax: -6780
raguilfo@facstaff.wisc.edu

The objective of this project is to continue developing more efficient tools and methods addressing the "front-end" processes of large-scale DNA sequencing. Our specific aims are high-throughput purification and mapping of cosmid inserts, controlled fragmentation of random inserts, direct selection vectors for cloning and sequencing, high-throughput M13 clone isolations, and high-throughput template purifications.

An approach to multi-cosmid purifications was developed using a cell-harvester and binding to GF/C glass fiber filter-bottom microtiter plates. This method proved inadequate because the yields were low and the DNA was easily fragmented. In the last year we have started examining the use of triplex-affinity capture (TAC) for this purpose as applied to BACs, based on our previous success with TAC purification and restriction mapping of cosmids (1,2).

We initially proposed to control random fragmentation for shotgun cloning using CviJ1 and its methyltransferase. Instead, we are now exploring automating it by scaled-down nebulization and parallel processing.

We have made a vector, M13-102 (3,4, patented)), for facilitating construction and improving quality of M13 shotgun libraries. It allows direct selection of recombinants, dephosphorylation of inserts to reducing chimerics, contains universal primers for fluorescent sequencing, and a triplex sequence for easy TAC purification of linearized RF DNA. We also made a version of this vector, M13-100Z, which expressed the alpha-peptide of B-gal. Its utility is in flow cytometry based clone isolation. We continue to develop these vectors for multiple cloning sites, and insert flipping using in closing steps of large-scale sequencing projects.

We continue to develop high-throughput clone isolations by flow cytometric cell sorting. M13 or plasmid clones can theoretically be isolated at rates in microtiter wells at rates up to 2 per second using our present FacStar-Plus cytometer and collection assembly. Theoretical rates are much higher. This bypasses plating onto solid-media and any need for plaque/colony picking. We initially tried isolations after microencapsulation of cells in agarose gel microbeads, but with H/W and S/W improvements we can now distinguish positively selected transfected cells from background. Efficiency of sorting is very sensitive to detection efficiency. We continue to investigate different methods of florescence detection for various plasmid and M13 vector systems including fluorogenic substrates for B-gal, fluorescent-tagged antibodies to M13 or cell surface proteins, and green fluorescent protein as a reporter.

We have been developing a solid-phase filter plate method for M13 template purifications using carboxylated polystyrene beads (Bangs Labs, IN) for automating on the Hamilton 2200. It should process 96 samples in under 30 minutes and deliver 1-2 micrograms per sample for cycle-sequencing. This approach has proven superior to others we have tried with respect to amenability to automation (5,6).

Ancillary projects. We reported a method for direct fluorescence analysis of genetic polymorphisms using oligonucleotide arrays on glass supports (7), which spun off other projects including (a) enhanced discrimination by artificial mismatch hybridization (8), restriction hybridization ordering of shotgun clones, and restriction site indexing-PCR (RSI-PCR) (9, patent applied for). RSI-PCR is an alternative strategy to extra-long PCR which has application in large gap filling (>45kb) differential

gene expression analysis, RFLP and EST marker production, end-sequencing and others.

Our most significant findings are the following:

1. Improved direct selection M13 cloning vector
2. Rapid restriction mapping of cosmids using triple-helix affinity capture
3. High-throughput M13 template production using carboxylated beads
4. Sequencing of a cosmid encoding the Drosophila GABA receptor
5. Improved detection of sequencing clones by flow-cytometry
6. RSI-PCR, a strategy to obtain mapped and sequence-ready DNA directly from up to 0.5 kb regions of a complex genome using palindromic class II restriction enzymes; bypasses conventional cloning methodology (see previous section for applications).

DOE Grant No. DE-FG02-91ER61122.

References

1. Ji, H., Smith, L.M., and Guilfoyle, R.A. (1994) GATA 11, 43-47.
2. Ji, H., Francisco, T., Smith, L.M. and Guilfoyle, R.A. (1996) Genomics 31, 185-192.
3. Guilfoyle,R. and Smith, L.M. (1994) Nucleic Acids Res. 22, 100-107.
4. Chen, D., Johnson, A.F., Severin, J.M., Rank, D.R., Smith, L.M. and Guilfoyle, R.A. (1996) Gene 172, 53-57.
5. Kolner, D.E., Guilfoyle, R.A., and Smith, L. (1994) DNA Sequence 4, 253-257.
6. Johnson, A.F., Wang, R., Ji, H., Chen, D., Guilfoyle, R.A. and Smith, L.M. (1996) Anal Biochem 234, 83-95.
7. Guo, Z., Guilfoyle, R.A., Thiel, A.J., Wang, R. and Smith, L.M. (1994) Nucleic Acids Res, 22, 5456-5465.
8. Guo, Z., Liu, Q., and Smith, L.M. (submitted).
9. Guilfoyle, R.A., Guo, Z., Kroening, D., Leeck, C. and Smith, L.M.(submitted).

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

Lloyd M. Smith and Brian Chait[1]
Department of Chemistry; University of Wisconsin; Madison, WI 53706
608/263-2594, Fax: /265-6780, *smith@chem.wisc.edu*
[1]Rockefeller University; New York, NY 10021

Our mass spec research has focused primarily on the possibility of utilizing Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry (MALDI-MS) as an alternative method to conventional gel electrophoresis for DNA sequence analysis. In this approach, extension fragments generated by the Sanger sequencing reactions are separated by size and detected in the mass spectrometer in one step.

Our group has shown fragmentation to be a major factor limiting accessible mass range, sensitivity, and mass resolution in the analysis of DNA by MALDI-MS. This DNA fragmentation was shown to be strongly dependent on both the MALDI matrix and the nucleic acid sequence employed. Fragmentation is proposed to follow a pathway in which nucleobase protonation leads to cleavage of the N-glycosidic bond with base loss, followed by cleavage of the phosphodiester backbone. Modifications of the deoxyribose sugar ring by replacing the 2' hydrogen with more electron-withdrawing groups such as the hydroxyl or fluoro group were shown to stabilize the N-glycosidic bond, partially or completely blocking fragmentation at the modified nucleosides. The stabilization provided by these chemical modifications was also shown to expand the range of matrices useful for nucleic acid analysis, yielding in some cases greatly improved performance.

DOE Grant No. DE-FG02-91ER61130.

Relevant Publication

Zhu, L.; Parr, G. P.; Fitzgerald, M. C.; Nelson, C. M.; Smith, L. M. Oligodeoxynucleotide fragmentation in MALDI/TOF Mass spectrometry using 355 nm radiation. J. Am. Chem. Soc. 1995, 117, 6048-6056.

Analysis of Oligonucleotide Mixtures by Electrospray Ionization-Mass Spectrometry

Richard D. Smith, David C. Muddiman, James E. Bruce, and Harold R. Udseth
Environmental Molecular Sciences Laboratory; Pacific Northwest National Laboratory; Richland, WA 99352
509/376-0723, Fax: -5824, *rd_smith@pnl.gov*
http://www.emsl.pnl.gov:2080/docs/msd/fticr/advmasspec.html

This project aims 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. A second goal is develop mass spectrometric methods that can be effective utilized in post genomic research in broad areas of DNA characterization, such as with polymerase chain reaction to rapidly and accurately identify single base polymorphisms. ESI produces intact molecular ions from DNA fragments of different size and sequence with high efficiency [1]. Our aim is to determine ESI mass spectrometry conditions that are compatible with biological sample preparation to allow efficient ionization of DNA and allowing for the analysis of complex mixtures (e.g., Sanger sequencing ladder). We have developed a novel on-line microdialysis method at PNNL to remove salts, detergents, and buffers from such biological preparations as PCR and dideoxy sequencing mixtures. This has allowed for rapid and efficient desalting (e.g., of samples having 0.25 M NaCl) allowing ESI mass spectral analysis without the typically problematic Na-adducts observed. Oligonucleotide ions are typically produced from ESI with

a broad distribution of net charge states for each molecular species, 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 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 [2,3]. Several methods show promise for charge state reduction and results have been demonstrated for series of smaller oligonucleotides. We have recently demonstrated for the first time that PCR products can be rapidly detected using ESI-MS with significant improvements projected [4,5]. Finally, new mass spectrometric methods have been developed to provide the dynamic range expansion necessary for addressing DNA sequencing mixtures [6]. Our overall aim is to provide a foundation for the development of an overall approach to high speed sequencing (including the rapid and precise PCR product characterization) using cost effective high-throughput instrumentation.

DOE Contract No. DE-AC06-76RLO-1830.

References

[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, H.R. Udseth and R.D. Smith J. Am. Soc. Mass Spectrom., 7 (8) 697-706 (1996).

[4] "Analysis of Double-stranded Polymerase Chain Reaction Products from the Bacillus cereus Group by Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry" D.S. Wunschel, K.F. Fox, A. Fox, J.E. Bruce, D.C. Muddiman and R.D. Smith Rapid Commun. in Mass Spectrom., 10, 29-35 (1996).

[5] "Characterization of PCR Products From Bacilli Using Electrospray Ionization FTICR Mass Spectrometry", D. C. Muddiman, D. S. Wunschel, C. Liu, L. Pasa-Tolic, K. F. Fox, A. Fox, G. A. Anderson, and R. D. Smith, Anal. Chem., 68, 3705-3712 (1996).

[6] "Colored Noise Waveforms and Quadrupole Excitation for the Dynamic Range Expansion in Fourier Transform Ion Cyclotron Resonance Mass Spectrometry", J. E. Bruce, G. A. Anderson and R. D. Smith, Anal. Chem., 68, 534-541 (1996).

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

Richard D. Smith, David C. Muddiman, S. A. Hofstadler, and J. E. Bruce
Environmental Molecular Sciences Laboratory; Pacific Northwest National Laboratory; Richland, WA 99352
509/376-0723, Fax: -5824, *rd_smith@pnl.gov*
http://www.emsl.pnl.gov:2080/docs/msd/fticr/advmasspec.html

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), are 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) are then 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 new FTICR detection methodologies.

Initial efforts under 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 [1-6]. We have shown that large multiply-charged individual ions of both single and double-stranded DNA anions can also be efficiently trapped in an 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. One approach is to 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. Instrumentation optimized for these purposes is currently being introduced and promises to greatly advance the methodology. 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).

DOE Contract No. DE-AC06-76RLO-1830.

(abstract continued)

References

[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 (108 MDa) Ions by Electrospray Ionization Fourier Transform Ion Cyclotron Resonance 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).

[6] "Accurate Molecular Weight Determination of Plasmid DNA Using Mass Spectrometry", X. Cheng, D. G. Camp II, Q. Wu, R. Bakhtiar, D. L. Springer, B.J. Morris, J. E. Bruce, G. A. Anderson, C. G. Edmonds and R. D. Smith, Nucleic Acid Res., 24, 2183-2189 (1996).

Characterization and Modification of DNA Polymerases for Use in DNA Sequencing

Stanley Tabor
Harvard University; Boston, MA 02115-5730
617/432-3128, Fax: -3362, *tabor@bcmp.med.harvard.edu*
http://sbweb.med.harvard.edu/~bcmp
http://sbweb.med.harvard.edu/~bcmp/tabor.html

Our studies are directed towards improving the properties of DNA polymerases for use in DNA sequencing. The primary focus is understanding the mechanism by which DNA polymerases discriminate against nucleotide analogs, and the mechanism by which they incorporate nucleotides processively without dissociating from the DNA template.

We are comparing three DNA polymerases that have been used extensively for DNA sequencing; *E. coli* DNA polymerase I, T7 DNA polymerase, and Taq DNA polymerase. These are related to one another, and this homology has been exploited to construct active site hybrids that have been used to determine the structural basis for differences in their activities. Specifically, the hybrids have been used (1) to determine why *E. coli* DNA polymerase I and Taq DNA polymerase discriminate strongly against dideoxynucleotides, and (2) to understand how T7 DNA polymerase interacts with its processivity factor, thioredoxin, to confer high processivity.

Based on these studies, we have been able to modify Taq DNA polymerase and *E. coli* DNA polymerase I to make them incorporate dideoxynucleotides much more efficiently, and to have increased processivity in the presence of thioredoxin. The ability to incorporate dideoxynucleotides efficiently greatly improves the uniformity of band intensities on a DNA sequencing gel, thereby increasing the accuracy of the DNA sequence obtained. In addition, the efficient use of dideoxynucleotides reduces the amount of these analogs required for DNA sequencing, an important issue when using fluorescently modified dideoxy terminators. In an approach that complements these studies, we, in collaboration with Dr. Thomas Ellenberger (Harvard Medical School), are determining the crystal structure of T7 DNA polymerase in a complex with thioredoxin and a primer-template. Knowledge of this structure will allow the rationale design of specific mutations that will enable DNA polymerases to incorporate other analogs useful for DNA sequencing more efficiently, such as those with fluorescent moieties on the bases.

DOE Grant No. DE-FG02-96ER62251.

Relevant Publication

Tabor, S., and Richardson, C. C. (1995). A single residue in DNA polymerases of the *Escherichia coli* DNA polymerase I family is critical for distinguishing between deoxy-and dideoxyribonucleotides. Proc. Natl. Acad. Sci. U.S.A. 92, 6339-6343.

Bedford, E., Tabor, S. and Richardson, C. C. (1997). The thioredoxin binding domain of bacteriophage T7 DNA polymerase confers processivity on *Escherichia coli* DNA polymerase I. Proc. Natl. Acad. Sci. U.S.A. 94, 479-484.

Modular Primers for DNA Sequencing

Mugasimangalam Raja,[1,2] Dina Sonkin,[2] Lev Lvovsky,[2] and **Levy Ulanovsky**[1,2]
[1]Center for Mechanistic Biology and Biotechnology; Argonne National Laboratory, Argonne, IL 60439-4833
Ulanovsky: 630/252-3940; Fax: -3387, *levy@anl.gov*
[2]Dept. of Structural Biology; Weizmann Institute of Science; Rehovot 76100, Israel

We are developing molecular approaches to DNA sequencing enabling primer walking without the step of chemical synthesis of oligonucleotide primers between the walks. One such approach involves "modular primers" described earlier, consisting of 5-mers, 6-mers or 7-mers (selected from a presynthesized library), annealing to the template contiguously with each other. Another approach, that we have termed DENS (Differential Extension with Nucleotide Subsets), works by selectively extending a short primer, making it a long one at the intended site only. DENS starts with a limited initial extension of the primer (at 20-30 C) in the presence of only 2 out of the 4 possible dNTPs. The primer is extended by 6-9 bases or longer at the intended priming site, which is deliberately selected, (as is the two-dNTP set), to maximize the extension length. The subsequent sequencing/termination reaction at 60-65 C then accepts the extended primer at the intended site, but not at alternative sites, where the initial extension

(if any) is generally much shorter. DENS allows the use of primers as long as 8-mers (degenerate in 2 positions) which prime much more strongly than modular primers involving 5-7 mers and which (unlike the latter) can be used with thermostable polymerases, thus allowing cycle-sequencing with dye-terminators for Taq, as well as making double-stranded DNA sequencing more robust.

These technologies are expected to speed up genome sequencing in more than one way:

a) Reduction in redundancy would result from more efficient and rapid closure of even long gaps which are currently avoided at the price of 7-to 9-fold redundancy in shotgun. Instantly available primers would also improve the quality of sequencing. Stretches of sequence that have too low confidence level (high suspected error rate) can be resequenced without synthesizing new oligos and without growing any new subclones.

b) Further down the road, the completion of the automation of the closed cycle of primer walking will be made possible via the elimination of the need to synthesize the walking primers. Combined with the capillary sequencers, the instant availability of the walking primers should reduce the time per walking cycle from 2-3 days now to about 1.5-2.0 hours, an improvement in speed by a factor of 20-50.

c) The closed-end automation would minimize both the labor cost and human errors. As primer walking has minimal, if any, front-end and back-end bottlenecks inherent to shotgun, the cost of sequencing would be essentially that of reagents, 5 cents/base or less.

DOE Grant No. DE-FG02-94ER61831.

Time-of-Flight Mass Spectroscopy of DNA for Rapid Sequence

Peter Williams, Chau-Wen Chou, David Dogruel, Jennifer Krone, Kathy Lewis, and Randall Nelson
Department of Chemistry and Biochemistry; Arizona State University; Tempe, AZ 85287
602/965-4107, Fax: -2747, *pw@asu.edu*

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 ambiguous 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.

DOE Grant No. DE-FG02-91ER61127.

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, and Eliza Fung
Ames Laboratory and Department of Chemistry; Iowa State University; Ames, IA 50011
515/294-8062, Fax: -0266, *yeung@ameslab.gov*

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 indexmatching 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.

Relevant Publications

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).

Resolving Proteins Bound to Individual DNA Molecules

David Allison, Bruce Warmack, Mitch Doktycz, Tom Thundat, and Peter Hoyt
Molecular Imaging Group; Health Sciences Research Division; Oak Ridge National Laboratory; Oak Ridge, TN 37831-6123
Allison: 423/574-6199, Fax: -6210, *allisondp@ornl.gov*
Warmack: 423/574-6202, Fax: -6210, *rjw@ornl.gov*

We have precisely located sequence specific proteins bound to individual DNA molecules by direct AFM imaging. Using a mutant *Eco*R I endonuclease that site-specifically binds but doesn't cleave DNA, bound enzyme has been imaged and located, with an accuracy of ±1%, on well characterized plasmids and bacteriophage lambda DNA (48 kb). Cosmids have been mapped and, by incorporating methods for anchoring molecules to surfaces and straightening to prevent molecular entanglement, BAC-sized clones could be analyzed.

This direct imaging approach could be rapidly developed to locate other sequence-specific proteins on genomic clones. Enzymatic proteins, involved in identifying and repairing damaged or mutated regions on DNA molecules, could be imaged bound to lesion sites. Transcription factor proteins that identify gene-start regions and other regulatory proteins that modulate the expression of genes by binding to specific control sequences on DNA molecules could be precisely located on intact cloned DNAs.

Conventional gel-based techniques for identifying site-specific protein binding sites must rely upon fragment analysis for identifying restriction enzyme sites, or, for non-cutting proteins, upon gel-shift methods that can only address small DNA fragments. Conversely, AFM imaging is a general approach that is applicable to the analysis of all site-specific DNA protein interactions on large-insert clones. This technique could be developed for high-throughput analysis, can be accomplished by technicians, uses readily available relatively inexpensive instrumentation, and should be a technology fully transferable to most laboratories.

DOE Contract No. DE-AC05-840R21400.

*Improved Cell Electrotransformation by Macromolecules

Alexandre S. Boitsov, Boris V. Oskin, Anton O. Reshetin, and Stepan A. Boitsov
Department of Biophysics; St.Petersburg State Technical University; 195251 St. Petersburg, Russia
+7-812/277-5959, Fax: /247-2088 or /534-3314, *sasha@bioph.hop.stu.neva.ru*

Our work for 1996 and 1997 will include the following:

1. Comparative study of the kinetics of entry of DNA of different molecular forms into E.coli cells DH10B/r and DH5a during electrotransformation. Study of the optimal regimes of cell-wall permeabilization for the DH10B/r cells.

2. Study of the efficiency of BAC cloning in DH10B/r cells using new electrotransformation method. Optimization of the procedure for DH10B/r cells.

3. Modernization of the electronic equipment in accordance with results of the biological experiments. To expand the studies, we need to extend the capability of the instrumentation to increase its flexibility and to improve the accuracy and reproducibility of the electric fields we generate by incorporating electronic components with higher tolerances.

DOE Grant No. OR00033-93CIS015.

Overcoming Genome Mapping Bottlenecks

Charles R. Cantor
Center for Advanced Biotechnology; Boston University; Boston MA 02215
617/353-8500, Fax: 8501, *crc@enga.bu.edu*
http://eng.bu.edu/CAB

Most traditional DNA analysis is done based on fractionation of DNA by length. We have, instead, begun to explore the use of DNA sequences as capture and detection methods to expedite a number of procedures in genome analysis.

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 restriction 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

*Projects designated by an asterisk received small emergency grants following December 1992 site reviews by David Galas (formerly DOE Office of Health and Environmental Research, which was renamed Office of Biological and Environmental Research in 1997), Raymond Gesteland (University of Utah), and Elbert Branscomb (Lawrence Livermore National Laboratory).

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.

In other work we are using enzyme-enhanced sequencing by hybridization (SBH) as a device for the rapid preparation of DNA samples for mass spectrometry. For example, partially duplex DNA probes can capture and generate sequence ladders from any arbitrary DNA sequence. Current MALDI protocols allow sequence to be read to lengths of 50 to 60 bases. While this is probably insufficient for most de novo DNA sequencing, it is an extremely promising approach for comparative or diagnostic DNA sequencing.

DOE Grant No. DE-FG02-93ER61609.

Preparation of PAC Libraries

Joe Catanese, Baohui Zhao, Eirik Frengen, Chenyan Wu, Xiaoping Guan, Chira Chen, Eugenia Pietrzak, Panayotis A. Ioannou,[1] Julie Korenberg,[2] Joel Jessee,[3] and **Pieter J. de Jong**
Department of Human Genetics; Roswell Park Cancer Institute; Buffalo, NY 14263
de Jong: 716/845-3168, Fax: -8849
pieter@dejong.med.buffalo.edu
http://bacpac.med.buffalo.edu
[1]The Cyprus Institute of Neurology and Genetics; Nicosia, Cyprus
[2]Cedars Sinai Medical Center; Los Angeles, CA 90048
[3]Life Technologies, Gaithersburg, MD 20898

Recently, we have developed procedures for the cloning of large DNA fragments using a bacteriophage P1 derived vector, pCYPAC1 (loannou 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 noncommercial 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.

DOE Grant No. DE-FG02-94ER61883 and NIH Grant No. 1R01RG01165.

Development of Affinity Technology for Isolating Individual Human Chromosomes by Third-Strand Binding

Jacques R. Fresco and Marion D. Johnson III
Department of Molecular Biology; Princeton University; Princeton, NJ 08544-1011
609/258-3927, Fax: -6730
esteckman@molbiol.princeton.edu
http://molbiol.princeton.edu

Prior to the onset of this grant, solution conditions had been developed for binding a 17-residue third strand oligodeoxyribonucleotide probe to a specific human chromosome (HC) 17 multicopy alpha satellite target sequence cloned into DNA vectors of varying size up to 50 kb. Binding was shown to be both highly efficient and specific. Moreover, initial experiments with fluorescent-labeled third strands and human lymphocyte metaphase spreads and interphase nuclei proved similarly successful. During the current research period, the technology for such third strand-based cytogenetic examination, i.e., Triplex *In Situ* Hybridization or TISH, of such spreads was perfected, so that it is now a highly reproducible method. Comparison of spreads of different individuals by TISH and FISH analysis has provided a new basis for detecting alpha satellite DNA polymorphisms, the basis of which requires further investigation.

This year work also commenced on the development of comparable probes specific for alpha satellite sequences in HC-X, 11, and 16. The work with HC-X has reached the stage where we are ready to test the probe for TISH-based cytogenetic analysis. Solution studies of the interaction of the probes designed for HC-11 and HC-16 alpha satellite targets are following the well-established path we employed for HC-17 and HC-X. With the expectation of success in these cases during the coming year, the way should be clear for the development and application of comparable probes for alpha satellite sequences of any other human chromosomes that may be of interest, and possibly of other eukaryotic species.

Meanwhile, we have begun to turn our attention to two other goals, one being the exploitation of our probes for the isolation of individual human chromosomes by affinity purification, as we originally proposed. The other goal is to exploit our probes as aids in flow sorting human chromosomes, a direction of work we expect to pursue in collaboration with the Los Alamos National Laboratory, just as soon as they indicate a readiness to do so. Finally, we have begun to evaluate the possibility of using third-strand binding fluorescent probes for detection of single copy genes by means of photon counting, a goal which we plan to undertake with our colleague Robert Austin of our Physics Department.

DOE Grant No. DE-FG02-96ER622202.

Chromosome Region-Specific Libraries for Human Genome Analysis

Fa-Ten Kao
Eleanor Roosevelt Institute for Cancer Research; Denver, CO 80206
303/333-4515, Fax: -8423, *kao@eri.uchsc.edu*

The objective of this project is to construct and characterize chromosome region-specific libraries as resources for genome analysis. We have used our chromosome microdissection and MboI linker-adaptor technique (PNAS 88, 1844, 1991) to construct region-specific libraries for human chromosome 2 and other chromosomes. The libraries have been critically evaluated for high quality, including insert size, proportion of unique vs repetitive sequence microclones, percentage of microclones derived from dissected region, etc.

We have constructed and characterized 11 region-specific libraries for the entire human chromosome 2 (the second largest human chromosome with 243 Mb of DNA), including 4 libraries for the short arm and 6 libraries for the long arm, plus a library for the centromere region. The libraries are large, containing hundreds of thousands of microclones in plasmid vector pUC19, with a mean insert size of 200 bp. About 40-60% of the microclones contain unique sequences, and between 70-90% 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 used in isolating other clones with large inserts (like YAC, BAC, PAC, P1 or cosmid) for contig assembly. These libraries have been used successfully for high resolution physical mapping and for positional cloning of disease-related genes assigned to these regions, e.g. the cloning of the gene for hereditary nonpolypsis colorectal cancer (Cell 75, 1215, 1993).

For each library, we have established a plasmid sub-library containing at least 20,000 independent microclones. These sub-libraries have been deposited to ATCC for permanent maintenance and general distribution. The ATCC Repository numbers for these libraries are: #87188 for 2P1 library

(region 2p23-p25, comprising 25 Mb); #87189 for 2P2 library (2p21-p23, 28 Mb); #87103 for 2P3 library (2p14-p16, 22 Mb); #87104 for 2P4 library (2p11-p13, 28 Mb); #77419 for 2Q1 library (2q35-q37, 28 Mb); #87308 for 2Q2 library (2q33-q35, 24 Mb); #87309 for 2Q3 library (2q31-q32, 26 Mb); #87310 for 2Q4 library (2q23-q24, 19 Mb); #87409 for 2Q5 library (2q21-q22, 23 Mb); #87410 for 2Q6 lbrary (2q11-q14, 31 Mb); and #87411 for 2CEN library (2p11.1-q11.1, 4 Mb). Details of these libraries have been described: Hum. Genet. 93, 557, 1994 (for 2P1 library); Cytogenet. Cell Genet. 68, 17, 1995 (for 2P2 library); Somat. Cell Mol. Genet. 20, 353, 1994 (for 2P3 library); Somat. Cell Mol. Genet. 20, 133, 1994 (for 2P4 library); Genomics 14, 769, 1992 (for 2Q1 library; Somat. Cell Mol. Genet. 21, 335, 1995 (for 2Q2, 2Q3 & 2Q4 libraries); Somat. Cell Mol. Genet. 22, 57, 1996 (for 2Q5, 2Q6 & 2CEN libraries).

Region-specific libraries and short insert microclones for chromosome 2 are particularly useful resources for its eventual sequencing because this chromosome is less exploited and detailed mapping information is lacking. We have also constructed 3 region-specific libraries for the entire chromosome 18 using similar methodologies, including 18P library (18p11.32-p11.1, 22 Mb); 18Q1 library (18q11.1-q12.3, 25 Mb); and 18Q2 library (18q21.1-q23, 34 Mb). Details of these libraries have been described (Somat. Cell Mol. Genet. 22, 191-199, 1996).

DOE Grant No. DE-FG03-94ER61819.

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

V.L. Karpov, O.V. Preobrazhenskaya, S.V. Belikov, and D.E. Kamashev
Engelhardt Institute of Molecular Biology; Russian Academy of Sciences; Moscow 17984, Russia
Fax: +7-095/135-1405, *karpov@genom-II.eimb.rssi.ru*

In 1995-1996 we continued to map and identify nonhistone proteins binding at loci along the yeast chromosome. Using DNA-protein crosslinking in vivo, we detected two polypeptides that probably correspond to core subunits of yeast RNA-polymerase II in the coding region of the transketolase gene (TKL2). Several nonhistone proteins were detected that bind to the upstream region of TKL2 and to an intergenic spacer between calmodulin (CMD1) and mannosyl transferase (ALG1) genes. The apparent molecular weight of these proteins was estimated. We also developed a new method to synthesize strand-specific probes.

Using DNA-protein crosslinking in vitro, we found the amino acid residues of the Lac-repressor that interacts with DNA. Only Lys-33 crosslinks with the Lac-operator in the specific complex.

In addition to Lys-33, the N-terminal end of the protein also crosslinks in a nonspecific complex. Our results demonstrate that, in the presence of an inducer, the repressor's N-termini crosslink to the operator's outermost nucleotides. We suggest that binding of an inducer changes the orientation of the DNA-binding domain of the Lac repressor to the opposite of that found for the specific complex.

We plan to use a new method to increase resolution and thus identify amino acids and nucleotides that participate in DNA-protein recognition. The mechanisms of transcription regulation of some yeast genes will thus be further elucidated. Our approaches are based on DNA-protein crosslinking. Detailed analysis will be done for specific and nonspecific complexes, in the presence and absence of inducers. This will allow us to make some conclusions about possible conformational rearrangements in DNA-protein complexes during gene activation at the protein's DNA-binding domains.

DOE Grant No. OR00033-93C1S007.

References

1. Papatsenko D.A., Belikov S.V., Preobrazhenskaya O.V., and Karpov V.L. Two-dimensional gels and hybrydization for studying DNA-protein contacts by crosslinking // Methods in Molecular and Cellular Biology. 1995. V. 5, No 3. P.171-177.
2. Kamashev D., Esipova N.G., Ebralidse K., and Mirzabekov, A.D. Mechanism of lac repressor switch-off: Orientation of lac repressor DNA-binding domain is reversed upon inducer binding //FEBS Lett. 1995. V.375. P.27-30
3. Papatsenko D.A., Priporova I.V., Belikov S.V., and Karpov, V.L. Mapping of DNA-binding proteins along yeast genome by UV-induced DNA-protein crosslinking.// FEBS Letters, 1996, 381, 103-105.
4. Belikov S.V., Papatsenko D.A., and Karpov V.L. A method to synthesize strand-specific probes. //Anal.Biochemistry, 1996, 240,152-154.

A PAC/BAC Data Resource for Sequencing Complex Regions of the Human Genome: A 2-Year Pilot Study

Julie R. Korenberg
Cedars Sinai Medical Center; University of California; Los Angeles, CA 90048-1869
310/855-7627, Fax: /652-8010
jkorenberg@mailgate.csmc.edu

While the complete sequencing the human genome at 99.99% accuracy is an immediate goal of the Human Genome Project, a serious technical deficiency remains the ability to rapidly and efficiently construct sequence ready maps as sequencing templates. This is particularly problematic in regions with unusual genome structure. An understanding of these troublesome regions prior to genome-wide sequencing will provide quality assurance as well as reliable sequencing strategies in these regions.

This proposal will generate a "whole genome" data resource to enable rapid and reliable sequencing of genomic DNA by the definition and characterization of the more than 52 regions of high homology now known to be distributed within unrelated genomic regions and cloned in BACs and PACs. To do this, we will:

1. Define regions of true homology in the human genome by characterizing subsets of the 4,700 BAC/PACs that generate multiple hybridization signals using fluorescence in situ hybridization (FISH). Of the 1,200 sites of multiple signals, more than 52 regions contain repeats as defined by 600 BAC/PACs. The chimerism rate, multiple clone wells, and chromosome of origin will be defined by re-streaking each clone, followed by fingerprint, FISH and PCR-based end-sequence analyses on hybrid panels and radiation hybrids.

Data will be shared with large sequencing efforts, deposited in the 4D database, available with annotation on ftp server and through GDB.

2. Generate contigs of BACs and PACs in regions of complex genome organization. Using STS, EST analyses, fingerprinting, BAC/PAC to BAC/PAC Southerns, end sequence walking in 3.5-20X libraries, and metaphase/interphase FISH, contigs will be seeded in 2-5 of the regions of known genome complexity, each of which is estimated as 2-5 Mb. These data will be used to evaluate and provide independent quality assurance of the STS and Radiation hybrid, and genetic maps in these regions. The most significant of these include 1p36/1q; 2p/q; multiple sites; 8p23 and 8 further sites; 9p/q.

3. Define additional regions of complex genomic structure. Library screening using known members of multiple member retro-transposon and other known repeated sequences defined by the ncbi database, followed by FISH analyzes to determine structure and potential large regions of associated homologies.

Collaboration with other genome and sequencing centers will provide quality control in the generation of sequence-ready maps for sequencing templates.

We believe that this effort is important since 1) it will provide a critical mapping tool necessary for the generation of sequence ready maps; 2) if initiated now, the problem areas could be delineated before scale ups to full production occur in major genome centers; 3) represents a modest cost such that the cost of these data would comprise only a small fraction of the cost of the entire genome sequence and would vastly decrease the cost of sequencing errors 4) and could be completed in a, short time (2 to 3 years) so as to be of maximum benefit to sequencing centers. The Principal Investigator in this project is ideally suited for this effort because the group has developed the technology and initiated FISH and genome analyses of over 4000 clones.

We believe that this project represents a critical and timely effort to enable rapid and cost effective human genome sequencing.

Subcontract under Glen Evans' DOE Grant No. DE-FC03-96ER62294.

Mapping and Sequencing of the Human X Chromosome

D. L. Nelson, E.E. Eichler, B.A. Firulli, Y. Gu, J. Wu, E. Brundage, A.C. Chinault, M. Graves, A. Arenson, R. Smith, E.J. Roth, H.Y. Zoghbi, Y. Shen, M.A. Wentland, D.M. Muzny. J. Lu, K Timms, M. Metzger, and R.A. Gibbs
Department of Molecular and Human Genetics and Human Genome Center; Baylor College of Medicine; Houston, TX 77030
713/798-4787, Fax: -6370 or -5386, *nelson@bcm.tmc.edu*
http://www.bcm.tmc.edu/molgen

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. Over 50 YACs have been utilized as probes to the gridded cosmic 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. More recent emphasis has been placed on BAC clones as their identity for sequencing has been established. These have been identified using the usual methods.

Cosmid contig construction. Creation of long-range continuity in cosmids and BACs 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 ~700 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 16pl 1 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.

DOE Grant Nos. DE-FG05-92ER6l401 and DE-FG03-94ER61830 and NIH Grant No. 5P30 HG00210.

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

Olga Podgornaya, Ivan Lobov, Ivan Matveev, Dmitry Lukjanov, Natella Enukashvily, and Elena Bugaeva
Institute of Cytology; Russian Academy of Sciences; St. Petersburg 194064, Russia
Telephone and Fax: +7-812/520-9703
podg@ivm.stud.pu.ru

Repetitive sequences occupy the most part of the whole eukaryotic genome but up to the last few years there has not been much interest in their role. The situation changed when alpha-satellites in human and minor satellites in mouse became candidates for centromere function responsibility. A number of centromere-specific proteins are under investigation but none seems to distinguish centromeric functions of exact sequences among long arrays of tandemly repeated satellites. The proteins associated with that array are poorly known. We are trying to find out what proteins are involved in maintaining the heterochromatin structure of different types of repetitive sequences.

The major proportion of total genomic satellite DNA remains attached to the nuclear matrix (NM) after DNase1 and high salt treatment. We followed this association in various steps during NM preparation by in situ hybridization with the mouse satellite probe. Two mouse species were used -*M. musculus* and *M. spretus*. Both contain the same repertoire of satellite DNAs but in different amounts. In *M. musculus* the centromeric heterochromatin contains major satellite (MA) as the principal component. In *M. spretus* the minor satellite (MI) is predominant. To test DNA-binding activity of the proteins after chromatography of the soluble NM proteins on cationic and anionic ion-exchange columns, gel shift assays were performed with cloned dimer of MA and a trimer of MI. To produce antibodies, the DNA-protein complexes obtained from large-scale gel-shift assays were isolated and injected into a guinea pig.

The gel shift assay with column fractions from *M. musculus* NM and MA shows a ladder of complexes. The complexes could be competed out with an excess of MA DNA but not with the same amount of *E. coli* DNA. Antibodies from the immune serum caused a hypershift of the MA/NM protein complexes. Preimmune serum at the same dilution did not alter the mobility of the complexes. A combination of western and Southern blots allows us to conclude that a protein with a molecular weight of about 80 kD and some similarity to the intermediate filaments is responsible for the MA/NM interaction.

Specific DNA-binding activity to the MI has been tested after column fractionation of the *M. spretus* NM extract. A ladder of complexes can be competed out with an excess of unlabeled MI but not *E. coli* or MA DNA. MI contains the CENPB-box sequence, which is the binding site for the protein CENPB, one of the centromeric proteins. Fractions from the NM extract with MI-specific binding activity do not contain CENPB, as shown by western blotting with anti-CENPB antibodies.

The same kind of work is going on with human analogs of MA and MI sequences, using large clones of satellite and alpha-satellite DNA and nuclear matrices.

There are few satellite DNA-binding proteins isolated, none of them directly from the NM. Our long-term aim is to understand the role of these proteins in heterochromatin formation and in heterochromatin association with NM.

Extracts from hand-isolated nuclear envelopes from frog oocytes were tested for the specific DNA-binding activity to (T2G4)116. A fragment of *Tetrahymena* telomere from a YAC plasmid was used as a labelled probe in a gel-shift assay. The DNA-protein complexes from the assay were cut out and injected into a guinea pig. The antibodies (AB) obtained stained one protein with an m.w. of about 70 kD in the nuclear envelope of the oocyte, nothing in the inner part of the oocyte, and 70 kD and 120 kD in the frog liver nuclei. The immunofluorescent AB stained fine patches on the oocyte nuclear envelope and a number of intranuclei spots in the frog blood cells.

The electron-microscope immuno-gold technique showed that the protein is localized in the outer surface of the oocyte nuclear envelope in cup-like structures. DNA-binding activity to the same sequence has been tested and found in the mouse nuclear matrix extracts. The activity could be eluted from the DEAE52 ion exchange column in 0.15 NaCl. The activity could be competed out with the fragment itself but not with *E. coli* DNA in the same amounts. AB stained a 70-kD protein in active fractions after ion exchange chromatography. In nuclear matrix preparations, the AB recognized a 120-kD protein as well. The AB caused hypershift of the complexes on the gel shift assay. The AB has some affinity to the keratins. In the mouse cell culture 3T3 line the staining is intranuclei, with fine dots forming chains surrounding dark areas, which do not correspond to the nucleoli.

Similar results were observed when a mouse cell line was transformed with head-and tail-less human keratin constructs (Bader et al., 1991, *J Cell Biol* **115**:1293). These results suggest that the nuclear proteins detected with the AB may be natural analogs of this artificial keratin construct. The pattern of staining did not resemble the picture of telomere-specific staining. Possibly the protein recognized intragenomic (T2G4)2 sequence, which is present in 25% of murine GenBank sequences rather than telomere. We are going to do immunocytochemical investigations of frog and mouse development in order to determine the point when transcription of the 120- kD protein is initiated and the staining becomes intranuclear.

As a continuation of the previous project the multiple alignment of all the *Alu* sequences from GenBank is going on. We are also trying to obtain antibodies to the main *Alu*-binding proteins to find out how many proteins could be bound to *Alu* sequence.

DOE Grant No. OR00033-93C1S014.

*Protein-Binding DNA Sequences

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

POU domain of Oct-2 transcription factor binds octamer sequence ATGCAAAT and a number of degenerated sequences. It has been shown that POUs and POUh domains recognize left and right parts of the oct-sequence, respectively. The recognized sequences are partly overlapped in the native octamer. In the degenerated recognition sites these core sequences may be separated with a spacer up to four nucleotides. The obtained data changed our view on the number and structure of potential targets recognized on DNA by POU proteins.

Protein-DNA binding is realized due to interaction of a conservative amino acid residues with a DNA target. In POU proteins amino acid residues in positions 47 (Val), 50 (Cys) and 51 (Asn) of POUh domain are absolutely conservative. In order to examine a possible role of Val47 we substituted this residue by each of the 19 other amino acid residues and the interaction of the mutant proteins was investigated with homeospecific site and its variants (ATAANNN) and with oct sequence. It was shown that Ile47 mutant retains the affinity and specificity. Val replacement for Ser, Thr or His partially reduce the affinity.

Asn47 mutant sharply relax the specificity of protein-DNA recognition. Mutants at 47 position have much stronger effects on binding to homeospecific sites than to octamer motifs. Our data indicate that there is not a simple mono-letter code of protein/DNA recognition. It has been shown that this recognition is determined not only by the nature of the radicals involved in the contact but also by the structure of DNA binding domain as a whole and probably by cooperative interaction of POUs and POUh domains.

Proposals for 1997. The role of Cys50 in POU domain/DNA recognition will be investigated. This residue is absolutely conservative in POU proteins but it is variable in relative homeo-proteins. Our preliminary data allow to suppose that residue at position 50 of POU homeodomain have a key role in discrimination between TAAT-like and octamer sequences. The role of the nuleotides flanking DNA target will be investigated.

DOE Grant No. OR00033-93CIS005.

Relevant Publications

1. Stepchenko A.G. (1994) Noncanonical oct-sequences are targets for mouse Oct-2B transcription factor. FEBS Letters, V.337, P.175-178.
2. Stepchenko A.G., Polanovsky O.L. (1996) Interaction of Oct proteins with DNA. Molecular Biology, V.30, P.296-302.
3. Stepchenko A.G., Luchina N.N., Polanovsky O.L. The role of conservative Val47 for POU homeodomain/DNA recognition. FEBS Letters, in press.

*Development of Intracellular Flow Karyotype Analysis

V.V. Zenin,[1] N.D. Aksenov,[1] A.N. Shatrova,[1] N.V. Klopov,[2] L.S. Cram,[3] and **A.I. Poletaev**
Engelhardt Institute of Molecular Biology; Russian Academy of Sciences; Moscow 117984, Russia
Poletaev: +7-095/135-9824, Fax: -1405
polet@polet.msk.su
[1]Institute of Cytology; Russian Academy of Sciences; St. Petersburg, Russia
[2]St. Petersburg Institute of Nuclear Physics; Gatchina, Russia
[3]Los Alamos National Laboratory; Los Alamos, NM 87545

Instrumentation for univariate fluorescent flow analysis of chromosome sets has been developed for human cells. A new method of cell preparation and intracellular staining of chromosome with different dyes was developed and improved. Cells suspension for flow analysis must satisfy the following requirements: minimal amount of free chromosomes and debris (dead cells, cell fragments etc.); chromosomes structure must be stabilized inside mitotic cells; chromosomes must be stained inside the cells up to saturation with the used dyes; chromosomes must be able to release from cells with minimal possible mechanical treatment. The method includes enzyme treatment (chymotrypsin), incubation with saponin and separation of prestained cells from debris on sucrose gradient. The developed protocol was tested and improved in the course of several months of work and allows us to obtain a well stained sample with a minimal amount of contaminates [2].

A special magnetic mixing/stirring device was constructed to perform cell membrane breaking. It was placed inside the flow chamber of a serial flow cytometer ATC-3000 equipped with additional electronic card for time-gated data acquisition [1]. The rupturing of prestained mitotic cells is performed by means of a small magnetic rod vibrating in an alternative magnetic field. The efficiency of mitotic cells breaking with electromagnetic cell breaking device was tested using different human cell lines[2,3].

The device works in a stepwise mode: a defined volume of sample is delivered to the breaking chamber for rupturing mitotic cell (cells) for a defined time period, followed by buffer wash to move the released chromosomes from the breaking chamber to the point of the analysis. The information about the chromosomes appearing at the point of analysis is accumulated in list mode files, making 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 [2,3]. In addition it's possible to determine the presence of extra chromosomes or loss of chromosome types. Thus this approach combines the high performance of flow cytometry (quantitation and high throughput) with the advantages of image analysis (cell to cell karyotype analysis and skills of 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 clinical cytogeneticists. The results now obtained are in good accordance with goals of the project formulated before [4].

DOE Grant No. OR00033-93CIS008.

References

[1]. V.V. Zenin, N.D. Aksenov, A.N. Shatrova, Y.V. Kravatsky, A. Kuznetzova, L.S. Cram, A.I. Poletaev. "Time-gated human chromosome flow analysis" XVII Congress of the International Society for Analytical Cytology, 1994, Lake Placid, USA, Cytometry Supplement 7, p. 68.
[2]. V.V. Zenin, N.D. Aksenov, A.N. Shatrova, Y.V. Kravatsky, A. Kuznetsova, L.S. Cram , A.I. Poletaev: "Time-gated flow analysis of human chromosomes"; DOE Human Genome Program, Contractor-Grantee Workshop IV, November 13-17, 1994; Santa Fe, New Mexico, p. 13.
[3]. V.V. Zenin, N.D. Aksenov, A.N. Shatrova, N.V. Klopov , L.S. Cram, A.I. Poletaev: "Cell by cell flow analysis of human chromosome sets"; DOE Human Genome Program, Contractor-Grantee Workshop V, January 28-February 1,1996; Santa Fe, New Mexico, p. 112.
[4]. Andrei I. Poletaev, Sergei I. Stepanov, Valeri V. Zenin, Nikolay Aksenov, Tatijana V. Nasedkina and Yuri V. Kravazky: "Development of Intracellular FlowKaryotype Analysis"; DOE Human Genome, 1993 Program Report, p.34-35.

Mapping and Sequencing with BACs and Fosmids

Ung-Jin Kim, Hiroaki Shizuya, and **Melvin I. Simon**
Division of Biology; California Institute of Technology; Pasadena, CA 91125
Kim: 818/395-4901, Fax: /796-7066, *ung@caltech.edu*
Simon: 818/395-3944, Fax: /796-7066
simonm@starbase1.caltech.edu
http://www.tree.caltech.edu

BACs and fosmids are stable, nonchimeric, and highly representative cloning systems. BACs maintain large-fragment genomic inserts (100 to 300 kb) that are easily prepared for most types of experiments, including DNA sequencing.

We have improved the methods for generating BACs and developed extensive BAC libraries. We have constructed human BAC libraries with more than 175,000 clones from male fibroblast and sperm, and a mouse BAC library with more than 200,000 clones. We are currently expanding human library with the aim of achieving total 50X coverage human genomic library using sperm samples from anonymous donors.

The BAC libraries provide resources to bridge the gap between genetic-cytogenetic information and detailed physical characteristics of genomic regions that include DNA sequence information. They also provide reliable tools for generating a high-resolution, integrated map on which a variety of information and resources are correlated. Using primarily the human BAC library constructed from fibroblasts, we have assembled a physical contig map of chromosome 22 [1]. First, the entire library was screened by most of the known chromosome 22-specific markers that include cDNA, anonymous STS markers, FISH-mapped cosmids and fosmids, YAC-Alu PCR products, FISH-mapped BACs, and flow-sorted chromosome 22 DNA. The positive clones have been assembled into contigs by means of the STS-contents or other markers assigned to BAC clones. Most of the contigs were confirmed by using a restriction fingerprinting scheme originally developed by Sulston and Coulson, and modified in our laboratory. Currently, the contigs cover over 80% of the chromosome arm. Various physical or genetic landmarks on this chromosome can now be precisely localized simply by assigning them to BACs or contigs on the map. Using BAC end sequence information from each of the chromosome 22-specific BACs, it is now possible to close the gaps efficiently by screening deeper BAC libraries with new probes specific to the ends of contigs.

The resulting BAC contig map is now serving as a road map for sequencing the chromosome. Chromosome 22-specific BAC clones have been distributed to our collaborators including The Sanger Center and Dr. Bruce Roe in University of Oklahoma, and many of the clones have already been sequenced. BAC end sequencing scheme[2] will play a crucial role toward the complete sequencing of chromosome 22, and we are currently sequencing the ends of these BACs directly using the miniprepped BAC DNA as templates.

DOE Grant No. DE-FG03-89ER60891.

References

[1] Kim et al. (1996) A Bacterial Artificial Chromosome-based framework contig map of human chromosome 22q. Proc. Natl. Acad. Sci. USA v93 (13): pp6297-6301.
[2] Venter, C., Smith, H.O., and Hood, L. (1996) Nature 381: pp364-366.

Towards a Globally Integrated, Sequence-Ready BAC Map of the Human Genome

Ung-Jin Kim, Hiroaki Shizuya, and **Melvin I. Simon**
Division of Biology; California Institute of Technology; Pasadena, CA 91125
Kim: 818/395-4901, Fax: /796-7066, *ung@caltech.edu*
Simon: 818/395-3944, Fax: /796-7066
simonm@starbase1.caltech.edu
http://www.tree.caltech.edu

BAC clones are ideal for genome analysis since they are non-chimeric, stably maintain large fragment genomic inserts (100-300 kb)[1], and it is easy to prepare BAC DNA samples for most types of experiments including DNA sequencing[2]. We have improved BAC cloning technique in the past years and constructed >20X human BAC libraries. As BACs are proving to be the most efficient reagents for large scale genomic sequencing, we intend to increase the depth of the library to 50X genomic equivalence. Using the ESTs, especially the Unigenes that have been chromosomally assigned by other means such as Radiation Hybrid mapping and YAC-based STS content mapping, we plan to organize the BAC library into a mapped resource. The resulting BAC-EST framework map will provide a high resolution EST (or gene) map and instant entry points for gene finding and large scale genomic sequencing. We also intend to determine the end sequences of the BAC inserts from a significant number of the clones (at least 350,000 clones or 15X genomic equivalence) within two years [3]. All the BAC-EST mapping data and BAC end sequences will be made available via public databases and WEB servers. The mapping data and end sequence information will dramatically facilitate the process of finding clones that extend the sequenced regions with minimal overlaps. Thus, the tagged BAC libraries will serve as a reliable and facile sequence-ready resource and an organizing tool to support and coordinate simultaneously multiple sequencing projects all over the genome.

DOE Grant No. DE-FC03-96ER62242.

References

[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, Valena Mancino, Cecilie Boysen, Hyung-Lyun Kang, Melvin I. Simon, and Hiroaki Shizuya. (1996) Genomics 34, 213-218.
[3] Venter, C, Smith, H.O., and Hood, L. (1996) Nature 381: pp364-366.

Generation of Normalized and Subtracted cDNA Libraries to Facilitate Gene Discovery

Marcelo Bento Soares, Maria de Fatima Bonaldo, Pierre Jelenc, and Susan Baumes
Department of Psychiatry; Columbia University; and The New York State Psychiatric Institute; New York, NY 10032
212/960-2313, Fax: /781-3577,
cuc@cuccfa.ccc.columbia.edu

Large-scale single-pass sequencing of cDNA clones randomly picked from libraries has proven quite powerful to identify genes and the use of normalized libraries in which the frequency of all cDNAs is within a narrow range has been shown to expedite the process by minimizing the redundant identification of the most prevalent mRNAs. In an

attempt to contribute to the ongoing gene discovery efforts, we have further optimized our original procedure for construction of normalized directionally cloned cDNA libraries[1] and we have successfully applied it to generate a number of human cDNA libraries from a variety of adult and fetal tissues [2]. To date we have constructed libraries from infant brain, fetal brain, adult brain, fetal liver-spleen, full-term and 8-9 week placentae, adult breast, retina, ovary tumor, melanocytes, parathyroid tumor, senescent fibroblasts, pineal glands, multiple sclerosis plaques, testis, B cells, fetal heart, fetal lung, 8-9 week fetuses and pregnant uterus. 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.

However, given the large scale nature of the ongoing sequencing efforts and the fact that a significant fraction of the human genes has been identified already, the discovery of novel cDNAs is becoming increasingly more challenging. In an effort to expedite this process further, in collaboration with Greg Lennon (LLNL) we have developed and applied subtractive hybridization strategies to eliminate pools of sequenced cDNAs from libraries yet to be surveyed. Briefly, single-stranded DNA obtained from pools of arrayed and sequence I.M.A.G.E. clones are used as templates for PCR amplification of cDNA inserts with flanking T7 and T3 primers. PCR amplification products are then used as drivers in hybridizations with normalized libraries in the form of single-stranded circles. The remaining single-stranded circles (subtracted library) are purified by hydroxyapatite chromatography, converted to double-stranded circles and electroporated into bacteria. Preliminary characterization of a subtracted fetal liver-spleen library indicates that the procedure is effective to enhance the representation of novel cDNAs.

In an effort to enhance the representation of full-length cDNAs in our libraries, as we strive towards our final objective of generating full-length normalized cDNA libraries, 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 has proven 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.

DOE Grant No. DE-FG02-91ER61233.

References

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

[2] Bonaldo, M.F., Lennon, G. and Soares, M.B. (1996). Normalization and subtraction: Two approaches to facilitate gene discovery. Genome Research 6, 791-806.

Mapping in Man-Mouse Homology Regions

Lisa Stubbs, Johannah Doyle, Ethan Carver, Mark Shannon, Joomyeong Kim, Linda Ashworth,[1] and Elbert Branscomb[1]
Biology Division; Oak Ridge National Laboratory; Oak Ridge, TN 37831
423/574-0854, Fax: -1283, *stubbsl@bioax1.bio.ornl.gov* or *stubbslj@ornl.gov*
[1]Human Genome Center; 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 widely in 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, upon elucidating mechanisms behind evolutionary rearrangements that distinguish chromosomes of mammalian species, and upon devising means of exploiting the relationships between the two genomes for the discovery and analysis of new genes and other functional units in mouse and man.

One of the larger contiguous blocks of mouse-human genomic homology includes the proximal portion of mouse chromosome 7 (Mmu7). 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 six internal rearrangements of gene sequence in mouse relative to man. One of these differences involve a small segment of H19ql3.4 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 six internal rearrangements, including two transpositions and four local

inversions, are clustered together at two sites; our data suggest that the rearrangements occurred in a coincident fashion, or were commonly associated with unstable DNA sequences at those sites. 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. The structure and conserved functions of genes within these and other clustered gene families located on H19 also represent an active line of interest to our group. More recently, we have extended mapping studies to include clustered gene families located in other chromosomal regions, and are working to define the borders of mouse-human syntenic segments on a broader, genome-wide scale.

DOE Contract No. DE-AC05-96OR22464 and Contract No. W-7405-ENG-48 with Lawrence Livermore National Laboratory.

Positional Cloning of Murine Genes

Lisa Stubbs, Cymbeline Culiat, Ethan Carver, Johannah Doyle, Laura Chittenden, Mitchell Walkowicz, Nestor Cacheiro, Greg Lennon,[1] Gary Wright,[2] Joe Rutledge,[3] Robert Nicholls,[4] and Walderico Generoso
Biology Division; Oak Ridge National Laboratory; Oak Ridge, TN 37831-8077
423/574-0854, Fax: -1283, *stubbsl@bioax1.bio.ornl.gov* or *stubbslj@ornl.gov*
[1]Human Genome Center; Lawrence Livermore National Laboratory; Livermore, CA 94550
[2]University of Texas Southwestern Medical Center at Dallas; Dallas, TX 75235
[3]Children's Hospital and Medical Center; University of Washington School of Medicine; Seattle, WA 98105
[4]Department of Genetics; Case Western Reserve University; Cleveland, Ohio

Chromosome rearrangements, notably deletions and translocations, have proved invaluable as tools in the mapping and molecular cloning of a acquired and inherited human diseases. Because balanced translocations are cytologically visible, and generally produce profound disturbances in both gene expression and DNA structure without necessarily disturbing the structure of multiple genes, this type of mutation provides an especially 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 anomalies, including obesity, polycystic kidneys, gastrointestinal disorders, 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 rapidly-obtained, crude-level mapping information available, we can readily assess possible relationships between newly arising mutant phenotypes and linked candidate genes or related diseases that map to homologous regions of the human genome. Using this approach, we have recently begun to define the map positions of several mutations. Mapping results have led us to the identification of candidate genes for two mutations: one associated with congenital deafness and predisposition to severe gastric ulcers, and another associated with late-onset obesity. So far, we have characterized only a fraction of the mouse strains that comprise this valuable, recently-generated mutant collection in detail. As a integral part of this program, we are actively exploring new strategies and integrating information, technology and resources derived from the Human Genome research effort, that promise to increase the efficiency of breakpoint mapping and cloning dramatically. The mutations are scattered widely throughout the mouse genome corresponding to a broad selection of human homology regions. As new breakpoints are mapped, 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 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.

In addition to the analysis of translocation mutants, we have also characterized other types of mouse mutations, including: (1) *tottering* and *leaner*, allelic mutations associated with ataxia and epilepsy in mice, and representing murine models for human diseases, familial hemiplaegic migraine and episodic ataxia, respectively; and (2) *jdf2*, a locus associated with mutations causing runting, neuromuscular tremors and male sterility which is located in a mouse region related to the Prader Willi-Angleman syndrome gene interval of human 15q11-q13. Both sets of mutations affect large, complex, and highly conserved genes, and provide important animal models for the exploration of the diverse roles their human counterparts may play in human disease. In concert with these gene cloning studies, we have been involved in exploring new means of exploiting mouse-human genomic conservation in the isolation of functionally-significant sequences from large cloned regions of human DNA. The methods we have developed hold great promise as an efficient tool for gene discovery in cloned genomic regions.

DOE Contract No. DE-AC05-96OR22464.

Human Artificial Episomal Chromosomes (HAECS) for Building Large Genomic Libraries

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

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 beta-lymphoblasts cells the HAEC DNA was stably established as 160-180 kb functional autonomously replicating episomes.[5,7] 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.

DOE Contract No. DE-FG05-91ER61135.

References

[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. Herpes viruses 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 beta-lymphocytes with Engineered Epstein-Barr Virus. Nature Medicine 1, 1303–1308 (1995).
[6] Sun, T.-Q., Livanos, E., & Vos, J.-M.H. Engineering a mini-herpesvirus as a general strategy to transduce up to 180 kb of functional self-replicating human mini-chromosomes. Gene Therapy 3, 1081–1088 (1996).
[7] Wang, S. & Vos, J.-M.H. An HSV/EBV based vector for High Efficient Gene Transfer to Human Cells in vitro/in vivo. J. Virol. 70, 8422–8430 (1996).

*Cosmid and cDNA Map of a Human Chromosome 13q14 Region Frequently Lost at B Cell Chronic Lymphocytic Leukemia

N.K. Yankovsky, B.I. Kapanadze, A.B. Semov, A.V. Baranova, and G.E. Sulimova
N.I. Vavilov Institute of General Genetics; Moscow 117809, Russia
+7-095/135-5363, Fax: -1289, *yankovsky@vigg.ru* and *bion@glas.apc.org* (send to both addresses)

We are mapping a human chromosome 13q14 region frequently lost at human blood malignancy cold B cell chronic lymphocytic leukemia (BCLL). The final goal of the project is to find putative oncosupressor gene lost in the region at BCLL. We have constructed a cosmid contig between D13S1168 and D13S25 loci in the region. The interval had been shown to be in the center of the BCLL associated deletions. The contig consists of more than 100 cosmids from LANL human chromosome 13 specific

library (LA13NC01). We estimated the distance between D13S1168 and D13S25 loci as about 540 kb. We are constructing a transcriptional map of the region. Seven different cDNA clones were found with two of the cosmid clones. All cosmids corresponding to the minimal tilling path between D13S1168 and D13S25 are being used as probes for screening new cDNA clones. I.M.A.G.E. Consortium (LLNL) cDNA clones assigned to 13q14 will be mapped against the cosmid contig. Mapped cDNA clones will be checked as candidate oncosupressor genes for BCLL.

BCM Server Core

Daniel Davison and Randall Smith
Baylor College of Medicine; Houston, TX 77030
713/798-3738, Fax: -3759, *davison@bcm.tmc.edu*
http://www.bcm.tmc.edu

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 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.

Grant Nos.: DOE, DE-FG03-9SER62097/A000; National Library of Medicine, R01-LM05792; National Science Foundation, BIR 91-11695; National Research Service Award, F32-HG00133-01; NIH, P30-HG00210 and R01-HG00973-01.

A Freely Sharable Database-Management System Designed for Use in Component-Based, Modular Genome Informatics Systems†

Steve Rozen,[1] Lincoln Stein,[1] and **Nathan Goodman**
The Jackson Laboratory; Bar Harbor, ME 04609
Goodman: 207/288-6158, Fax: -6078, *nat@jax.org*
[1]Whitehead Institute for Biomedical Research; Cambridge, MA 02139
http://goodman.jax.org
http://www.genome.wi.mit.edu/informatics/workflow

We are constructing a data-management component, built on top of commercial data-management products, tuned to the requirements of genome applications. 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 bimolecular sequences,
- provide built-in support for tracking laboratory work flows, and admit further extensions for other special-purpose types,
- allow core facilities to be readily extended to meet the diverse needs of biological applications

The core data manager is being constructed on top of Sybase, Oracle, and Informix Universal Server. The software is available free of charge and is freely redistributable.

We will be reporting progress on the core data manager's architecture and interface at the URLs above, and we solicit comments on its design.

DOE Grant No. DE-FG02-95ER62101.

†Originally called Database Management Research for the Human Genome Project, this project was initiated in 1995 at the Massachusetts Institute of Technology–Whitehead Institute.

*Projects designated by an asterisk received small emergency grants following December 1992 site reviews by David Galas (formerly DOE Office of Health and Environmental Research, which was renamed Office of Biological and Environmental Research in 1997), Raymond Gesteland (University of Utah), and Elbert Branscomb (Lawrence Livermore National Laboratory).

A Software Environment for Large-Scale Sequencing

Mark Graves
Department of Cell Biology; Baylor College of Medicine; Houston, TX 77030
713/798-8271, Fax: -3759; *mgraves@bcm.tmc.edu*
http://www.bcm.tmc.edu
http://stork.bcm.tmc.edu/gfp

Our approach is to implement software systems which manage primary laboratory sequence data and explore and annotate functional information in genome sequence and gene products.

Three software systems have been developed and are being used: two sequence data managers which use different sequence assembly packages, FAK and Phrap, and a series of analysis and annotation tools which are available via the Internet. In addition, we have developed a prototype application for data mining of sequence data as it is related to metabolic pathways.

Products of this project are the following:

1. GRM -a sequence reconstruction manager using the FAQ assembly engine (available since October 1995).

2. GFP -a sequence finishing support tool using the Phrap assembly engine (available since March 1996).

3. A series of gene recognition tools (available since early 1996).

4. A tool for visualizing metabolic pathways data and exploring sequence data related to metabolic pathways (prototype available since August 1996).

DOE Grant No. DE-FG03-94ER61618.

Generalized Hidden Markov Models for Genomic Sequence Analysis

David Haussler, Kevin Karplus,[1] and Richard Hughey[1]
Computer Science Department and [1]Computer Engineering Department; University of California; Santa Cruz, CA 95064
408/459 2105, Fax: -4829, *haussler@cse.ucsc.edu*
http://www.cse.ucsc.edu/research/compbio
http://www-hgc.lbl.gov/projects/genie.html

We have developed an integrated probabilistic method for locating genes in human DNA based on a generalized hidden Markov model (HMM). Each state of a generalized HMM represents a particular kind of region in DNA, such as an initial exon for a gene. The states are connected by transitions that model sites in DNA between adjacent regions, e.g. splice sites. In the full HMM, parametric statistical models are estimated for each of the states and transitions. Generalized HMMs allow a variety of choices for these models, such as neural networks, high order Markov models, etc. All that is required is that each model return a likelihood for the kind of region or transition it is supposed to model. These likelihoods are then combined by a dynamic programming method to compute the most likely annotation for a given DNA contig. Here the annotation simply consists of the locations of the transitions identified in the DNA, and the labeling of the regions between transitions with their corresponding states.

This method has been implemented in the genefinding program Genie, in collaboration with Frank Eeckman, Martin Reese and Nomi Harris at Lawrence Berkeley Labs. David Kulp, at UCSC, has been responsible for the core implementation. Martin Reese developed the splice site models, promoter models, and datasets. You can access Genie at the second www address given above, submit sequences, and have them annotated. Nomi Harris has written a display tool called Genotater that displays Genie's annotation along with the annotation of other genefinders, as well as the location of repetitive DNA, BLAST hits to the protein database, and other useful information. Papers and further information about Genie can be found at the first www address above. Since the ISMB '96 paper, Genie's exon models have been extended to explicitly incorporate BLAST and BLOCKS database hits into their probabilistic framework. This results in a substantial increase in gene predicting accuracy. Experimental results in tests using a standard set of annotated genes showed that Genie identified 95% of coding nucleotides correctly with a specificity of 88%, and 76% of exons were identified exactly.

DOE Grant No. DE-FG03-95ER62112.

Identification, Organization, and Analysis of Mammalian Repetitive DNA Information

Jerzy Jurka
Genetic Information Research Institute; Palo Alto, CA 94306
415/326-5588 Fax: -2001, *jurka@gnomic.stanford.edu*
http://charon.lpi.org

There are three major objectives in this project: organization of databases of mammalian repetitive sequences, development of specialized software for analysis of repetitive DNA, and sequence studies of new mammalian repeats.

Our approach is based on extensive usage of computer tools to investigate and organize publicly available sequence information. We also pursue collaborative research

with experimental laboratories. The results are widely disseminated via the internet, peer reviewed scientific publications and personal interactions. Our most recent research concentrates on mechanisms of retroposon integration in mammals (Jurka, J., PNAS, in press; Jurka, J and Klonowski, P., J. Mol. Evol. 43:685-689).

We continue to develop reference collections of mammalian repeats which became a worldwide resource for annotation and study of newly sequenced DNA. The reference collections are being revised annually as part of a larger database of repetitive DNA, called Repbase. The recent influx of sequence data to public databases created an unprecedented need for automatic annotation of known repetitive elements. We have designed and implemented a program for identification and elimination of repetitive DNA known as CENSOR.

Reference collections of mammalian repeats and the CENSOR program are available electronically (via anonymous ftp to ncbi.nih.gov; directory repository/repbase). CENSOR can also be run via electronic mail (mail "help" message to censor@charon.lpi.org).

DOE Grant No. DE-FG03-95ER62139.

*TRRD, GERD and COMPEL: Databases on Gene-Expression Regulation as a Tool for Analysis of Functional Genomic Sequences

A.E. Kel, O.A. Podkolodnaya, O.V. Kel, A.G. Romaschenko, E. Wingender,[1] G.C. Overton,[2] and **N.A. Kolchanov**
Institute of Cytology and Genetics; Novosibirsk, Russia
Kolchanov: +7-3832/353-335, Fax: -336 or /356-558,
kol@benpc.bionet.nsc.ru
http://transfac.gbf-braunschweig.de
[1]Gesellschaft für Biotechnologische Forschung; Braunschweig, Germany
[2]Department of Genetics; University of Pennsylvania School of Medicine; Philadelphia, PA 19104-6145

The database on transcription regulatory regions in eukaryotic genomes (TRRD) has been developed [1] (http://www.bionet.nsk.su/TRRD.html; ftp://ftp.bionet.nsk.su/pub/trrd/). The main principle of data representation in TRRD is modular structure and hierarchy of transcription regulatory regions. TRRD entry corresponds to a gene as entire unit. Information on gene regulation is provided (cell-cycle and cell type specificity, developmental stage-specificity, influence of various molecular signals on gene expression). TRRD database contains information about structural organization of gene transcription regulatory region. TRRD contains description of known promoters and enhancers in 5', 3' regions and in introns. Description of binding sites for transcription factors includes nucleotide sequence and precise location, name of factors that bind to the site, experimental evidences for the binding site revealing. We provide cross-references to TRANSFAC database [2] for both sites and factors as well as for genes. TRRD 3.3 release includes 340 vertebrate genes.

The Gene Expression Regulation Database (GERD) collects information on features of genes expression as well as information about gene transcription regulation. The current release of GERD contains 75 entries with information on expression regulation of genes expressed in hematopoietic tissues in the course of ontogenesis and blood cells differentiation. COMPEL database contains information about composite elements which are functional units essential for highly specific transcription regulation [3]. Direct interactions between transcription factors binding to their target sites within composite elements result in convergence of different signal transduction pathways. Nucleotide sequences and positions of composite elements, binding factors and types of their DNA binding domains, experimental evidence confirming synergistic or antagonistic action of factors are registered in COMPEL. Cross-references to TRANSFAC factors table are given. TRRD and COMPEL are provided by cross-references to each other. COMPEL 2.1 release includes 140 composite elements.

We have developed a software for analysis of transcription regulatory region structure. The CompSearch program is based on oligonucleotide weight matrix method. To collect sets of binding sites for the matrixes construction we have used TRANSFAC and TRRD databases. The CompSearch program takes into account the fine structure of experimentally confirmed NFATp/AP-1 composite elements collected in COMPEL (distances between binding sites in composite elements, their mutual orientation). By means of the program we have found potential composite elements of NFATp/AP-1 type in the regulatory regions of various cytokine genes. Analysis of composite elements could be the first approach to reveal specific patterns of transcription signals encoding regulatory potential of eukaryotic promoters.

References

1. Kel O.V., Romaschenko A.G., Kel A.E., Naumochkin A.N., Kolchanov N.A. Proceedings of the 28th Annual Hawaii International Conference on System Sciences [HICSS]. (1995), v.5, Biotechnology Computing, IEE Computer Society Press, Los Alamos, California, p. 42-51.
2. Wingender E., Dietze P., Karas H., and Knuppel R. TRANSFAC: a database on transcription factors and their DNA binding sites (1996). Nucl. Acids Res., 1996, v. 24, pp. 238-241.
3. Kel O.V., A.G. Romaschenko, A.E. Kel, E. Wingender, N.A. Kolchanov. A compilation of composite regulatory elements affecting gene transcription in vertebrates (1995). Nucl. Acids Res., v. 23, pp. 4097-4103.

(abstract continued)

Recent Publications

Kel, A., Kel, O., Ischenko, I., Kolchanov, N., Karas, H., Wingender, E. and Sklenar, H. (1996). TRRD and COMPEL databases on transcription linked to TRANSFAC as tools for analysis and recognition of regulatory sequences. Computer Science and Biology. Proceedings of the German Conference on Bioinformatics (GCB'96), R. Hofestadt, T. Lengauer, M. Löffler, D. Schomburg (eds.). University of Leipzig, Leipzig 1996, pp. 113-117.

Wingender, E., Kel, A. E., Kel, O. V., Karas, H., Heinemeyer, T., Dietze, P., Knueppel, R., Romaschenko, A. G. and Kolchanov, N. A. (1997). TRANSFAC, TRRD and COMPEL: Towards a federated database system on transcriptional regulation. Nucleic Acids Res., in press.

Ananko E.A., Ignatieva E.V., Kel A.E., Kolchanov N.A (1996). WWWTRRD: Hypertext information system on transcription regulation. Computer Science and Biology. Proceedings of the German Conference on Bioinformatics (GCB'96), R. Hofestadt, T. Lengauer, M. Löffler, D. Schomburg (eds.). University of Leipzig, Leipzig 1996, pp. 153-155.

A.E. Kel, O.V. Kel, O.V. Vishnevsky, M.P. Ponomarenko, I.V. Ischenko, H. Karas, N.A. Kolchanov, H. Sklenar, E. Wingender (1997). TRRD and COMPEL databases on transcription linked to TRANSFAC as tools for analysis and recognition of regulatory sequences. (1997) LECTURE NOTES IN COMPUTER SCIENCE, in press.

Holger Karas, Alexander Kel, Olga Kel, Nikolay Kolchanov, and Edgar Wingender (1997). Integrating knowledge on gene regulation by a federated database approach: TRANSFAC, TRRD and COMPEL. Jurnal Molekularnoy Biologii (Russian), in press.

Kel A.E., Kolchanov N.A., Kel O.V., Romaschenko A.G., Ananko E.A., Ignatyeva E.V., Merkulova T.I., Podkolodnaya O.A., Stepanenko I.L., Kochetov A.V., Kolpakov F.A., Podkolodniy N.L., Naumochkin A.A. (1997). TRRD: A database on transcription regulatory regions of eukaryotic genes. Jurnal Molekularnoy Biologii (Russian) in press.

O.V. Kel, A.E. Kel, A.G. Romaschenko, E. Wingender, N.A. Kolchanov (1997). Composite regulatory elements: classification and description in the COMPEL data base. Jurnal Molekularnoy Biologii (Russian), in press.

Data-Management Tools for Genomic Databases

Victor M. Markowitz and **I-Min A. Chen**
Information and Computing Sciences Division; Lawrence Berkeley National Laboratory; Berkeley, CA 94720
510/486-6835, Fax: -4004, *vmmarkowitz@lbl.gov*
http://gizmo.lbl.gov/opm.html

The Object-Protocol Model (OPM) data management tools provide facilities for constructing, maintaining, and exploring efficiently molecular biology databases. Molecular biology data are currently maintained in numerous molecular biology databases (MBDs), including large archival MBDs such as the Genome Database (GDB) at Johns Hopkins School of Medicine, the Genome Sequence Data Base (GSDB) at the National Center for Genome Resources, and the Protein Data Bank (PDB) at Brookhaven National Laboratory. Constructing, maintaining, and exploring MBDs entail complex and time-consuming processes.

The goal of the Object-Protocol Model (OPM) data management tools is to provide facilities for efficiently constructing, maintaining, and exploring MBDs, using application-specific constructs on top of commercial database management systems (DBMSs). The OPM tools will also provide facilities for reorganizing MBDs and for exploring seamlessly heterogenous MBDs. The OPM tools and documentation are available on the Web and are developed in close collaboration with groups maintaining MBDs, such as GDB, GSDB, and PDB.

Current work focuses on providing new facilities for constructing and exploring MBDs. The specific aims of this work are:

(1) Extend the OPM query language with additional constructs for expressing complex conditions, and enhance the OPM query optimizer for generating more efficient query plans.

(2) Develop enhanced OPM query interfaces supporting MBD-specific data types (e.g., protein data type) and operations (e.g., protein data display and 3D search), and assisting users in specifying and interpreting query results.

(3) Provide support for customizing MBD interfaces.

(4) Extend the OPM tools with facilities for managing permissions (object ownership) in MBDs, and for physical database design of relational MBDs, including specification of indexes, allocation of segments, and handling of redundant (denormalized) data.

(5) Develop OPM tools for constructing and maintaining multiple OPM views for both relational and non-relational (e.g., ASN.1, AceDB) MBDs. For a given MBD, these tools will allow customizing different OPM views for different groups of scientists. For heterogeneous MBDs, this tool will allow exploring them using common OPM interfaces.

(6) Develop tools for constructing OPM based multidatabase systems of heterogeneous MBDs and for exploring and manipulating data in these MBDs via OPM interfaces. As part of this effort, the OPM-based multidatabase system which consists currently of GDB 6.0 and GSDB 2.0, will be extended to include additional MBDs, primarily GSDB 2.2 (when it becomes available), PDB, and Genbank.

(7) Develop facilities for reorganizing OPM-based MBDs.The database reorganization tools will support automatic generation of procedures for reorganizing MBDs following restructuring (revision) of MBD schemas.

In the past year, the OPM data management tools have been extended in order to address specific requirements of developing MBDs such as GDB 6 and the new version of PDB.

The current version of the OPM data management tools (4.1) was released in June 1996 for Sun/OS, Sun/Solaris and SGI. The following OPM tools are available on the Web at http://gizmo.lbl.gov/opm.html:

(1) an editor for specifying OPM schemas;

(2) a translator of OPM schemas into relational database specifications and procedures;

(3) utilities for publishing OPM schemas in text (Latex), diagram (Postscript), and Html formats;

(4) a translator of OPM queries into SQL queries;

(5) a retrofitting tool for constructing OPM schemas (views) for existing relational genomic databases;

(6) a tool for constructing Web-based form interfaces to MBDs that have an OPM schema; this tool was developed by Stan Letovsky at Johns Hopkins School of Medicine, as part of a collaboration.

The OPM data management tools have been highly successful in developing new genomic databases, such as GDB 6 (released in January 1996; http://gdbgeneral.gdb.org/gdb/) and the relational version of PDB (http://terminator.pdb.bnl.gov:4148), and in constructing OPM views and interfaces for existing genomic databases such as GSDB 2.0. The OPM data management tools are currently used by over ten groups in USA and Europe. The research underlying these tools is described in several papers published in scientific journals and presented at database and genome conferences.

In the past year the OPM tools have been presented at database and bioinformatics conferences, including the International Symposium on Theoretical and Computational Genome Research, Heidelberg, Germany, March 1996, the Workshop on Structuring Biological Information, Heidelberg, Germany, March 1996, the Meeting on Genome Mapping and Sequencing, Cold Spring Harbor, May 1996, the International Sybase User Group Conference, May 1996, the Bioinformatics -Structure Conference, Jerusalem, November 1996, and the Pacific Symposium on Bioinformatics, January 1997.

The results of the research and development underlying the OPM tools work have been presented in papers published in proceedings of database and bioinformatics conferences; these papers are available at http://gizmo.lbl.gov/opm.html#Publications.

DOE Contract No. DE-AC03-76SF00098.

The Genome Topographer: System Design

S. Cozza, D. Cuddihy, R. Iwasaki, M. Mallison, C. Reed, J. Salit, A. Tracy, and **T. Marr**
Cold Spring Harbor Laboratory; Cold Spring Harbor, NY 11724
Marr: 516/367-8393, Fax: -8461, *marr@cshl.org* or *marr@cb.cshl.org*

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.

DOE Grant No. DE-FG02-91ER61190.

A Flexible Sequence Reconstructor for Large-Scale DNA Sequencing: A Customizable Software System for Fragment Assembly

Gene Myers and Susan Larson
Department of Computer Science; University of Arizona; Tucson, AZ 85721
602/621-6612, Fax: -4246, *gene@cs.arizona.edu*
http://www.cs.arizona.edu/faktory

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 multi-alignment of an assembly while consulting the raw waveforms. Special attention was given to optimizing the ergonomics of this time-intensive task.

DOE Grant No. DE-FG03-94ER61911.

The Role of Integrated Software and Databases in Genome Sequence Interpretation and Metabolic Reconstruction

Terry Gaasterland, Natalia Maltsev, **Ross Overbeek**, and Evgeni Selkov
Mathematics and Computer Science Division; Argonne National Laboratory; Argonne, IL 60439
630/252-4171, Fax: -5986, *gaasterl@mcs.anl.gov*
MAGPIE: *http://www.mcs.anl.gov/home/gaasterl/magpie.html*
WIT: *http://www.cme.msu.edu/WIT*

As scientists successfully sequence complete genomes, the issue of how to organize the large quantities of evolving sequence data becomes paramount. Through our work in comparative whole genome analysis (MAGPIE, Gaasterland) and metabolic reconstruction algorithms (WIT, Overbeek, Maltsev, and Selkov), we carry genome interpretation beyond the identification of gene products to customized views 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.[1,2] 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, and monitor the project sequence data for changes that initiate reanalysis.

In collaboration with Canada's Institute for Marine Biosciences and the Canadian Institute for Advanced Research, MAGPIE is being used to maintain and study comparative views of all open reading frames (ORFs) across fully sequenced genomes (currently 5), nearly completed genomes (currently 2) and 1 genome in progress (Sulfolobus solfataricus). Together, these genomes represent multiple archaeal and bacterial genomes and one eukaryotic genome. This analysis provides the necessary data to assign phylogenetic classifications to each ORF (e.g., "AE" for archaeal and eukaryotic). This data in turn provides the basis for validating and assessing functional annotations according to phylogenetic neighborhood (e.g., selecting the eukaryotic form of a biochemical function over a bacterial form for an "AE" ORF).[3]

Once an automated functional overview has been established, it remains to pinpoint the organisms' exact metabolic pathways and establish how they interact.To this end, the WIT (What Is There) system supports efforts to develop metabolic reconstructions. Such constructions, or models, are based on sequence data, clearly established biochemistry of specific organisms, understanding of the interdependencies of biochemical mechanisms. WIT thus offers a valuable tool for testing current hypotheses about microbial behavior. For example, a reconstruction may begin with a set of established enzymes (enzymes with strong similarities in identified coding regions to existing sequences for which the enzymatic function is known) and putative enzymes (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 (enzymes that have not been clearly detected, but that would be expected given the set of hypothesized pathways) and missing enzymes (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.

In addition to helping curators develop metabolic reconstructions, WIT lets users examine models curated by experts, follow connections between more than two thousand metabolic diagrams, and compare models (e.g., which of certain genes that are conserved among bacterial genomes are found in higher life). The objective is to set the stage for meaningful simulations of microbial behavior and thus to advance our understanding of microbial biochemistry and genetics.

DOE Contract No. W-31-109-Eng-38 (ANL FWP No. 60427).

References

[1] T. Gaasterland and C. Sensen, Fully Automated Genome Analysis that Reflects User Needs and Preferences --a Detailed Introduction to the MAGPIE System Architecture, Biochemie, 78(4), (accepted).
[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).

Database Transformations for Biological Applications

G. Christian Overton, **Susan B. Davidson**,[1] and **Peter Buneman**[1]
Department of Genetics and [1]Department of Computer and Information Science; University of Pennsylvania; Philadelphia, PA 19104
Overton: 215/573-3105, Fax: -3111, *coverton@cbil.humgen.upenn.edu*
Davidson: 215/898-3490, Fax: -0587, *susan@cis.upenn.edu*
Buneman: 215/898-7703, Fax: -0587,*peter@cis.upenn.edu*
http://agave.humgen.upenn.edu/cpl/cplhome.html
http://sdmc.iss.nus.sg/kleisli-stuff/MoreInfo.html

We have implemented a general-purpose query system, Kleisli, that provides access to a variety of "non-standard" 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, the CPL query language, 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 the 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 URL above).

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 data warehouses initialized from community databases. 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. We have examined in detail methods for 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.

Transformation of data resources, that is the structural reorganization of a data resource from one form to another, arises frequently in genome informatics. Examples include the creation of data warehouses and database evolution. 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, implementing and formally verifying transformations in a uniform way across a wide variety of different data models. Morphase is a prototype system for specifying transformations between data sources and targets in an intuitively appealing, declarative language based on Horn clause logic. Transformations specification in Morphase are translated into CPL and executed in the Kleisli system. The data-types underlying Morphase include arbitrarily nested records, sets, variants, lists and object identity, thus capturing the types common to most data formats relevant to genome informatics, including ASN.1 and ACE. Morphase can be connected to a wide variety of data sources simultaneously through Kleisli. 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. For example, we used it to specify a transformation between the Sanger Center's Chromosome 22 ACE database (ACE22DB) and a 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.

DOE Grant No. DE-FG02-94ER61923.

Relevant Publications

P. Buneman, S.B. Davidson, K. Hart, C. Overton and L. Wong,"A Data Transformation System for Biological Data Sources," in Proceedings of VLDB, Sept. 1995 (Zurich, Switzerland). Also available as Technical Report MS-CIS-95-10, University of Pennsylvania, March 1995.

S.B. Davidson, C. Overton and P. Buneman, "Challenges in Integrating Biological Data Sources," J. Computational Biology 2 (1995), pp 557-572.

A. Kosky, "Transforming Databases with Recursive Data Structures," PhD Thesis, December 1995.

S.B. Davidson and A. Kosky, "Effecting Database Transformations Using Morphase," Technical Report MS-CIS-96-05, University of Pennsylvania.

A. Kosky, S.B. Davidson and P. Buneman, "Semantics of Database Transformations," Technical Report MS-CIS-95-25, University of Pennsylvania, 1995.

K. Hart and L. Wong, "Pruning Nested Data Values Using Branch Expressions With Wildcards," In Abstracts of MIMBD, Cambridge, England, July 1995.

Las Vegas Algorithm for Gene Recognition: Suboptimal and Error-Tolerant Spliced Alignment

Sing Hoi Sze and **Pavel A. Pevzner**[1]
Departments of Computer Science and [1]Mathematics; University of Southern California; Los Angeles, CA 90089
Pevzner: 213/740-2407, Fax: -2424
ppevzner@hto.usc.edu
http://www-hto.usc.edu/software/procrustes

Recently, Gelfand, Mironov, and Pevzner (Proc. Natl. Acad. Sci. USA, 1996, 9061-9066) proposed a spliced alignment approach to gene recognition that provides 99% accurate recognition of human gene if a related mammalian protein is available. However, even 99% accurate gene predictions are insufficient for automated sequence annotation in large-scale sequencing projects and therefore have to be complemented by experimental gene verification. 100% accurate gene predictions would lead to a substantial reduction of experimental work on gene identification. Our goal is to develop an algorithm that either predicts an exon assembly with accuracy sufficient for sequence annotation or warns a biologist that the accuracy of a prediction is insufficient and further experimental work is required. We study suboptimal and error-tolerant spliced alignment problems as the first steps towards such an algorithm, and report an algorithm which provides 100% accurate recognition of human genes in 37% of cases (if a related mammalian protein is available). For 52% of genes, the algorithm predicts at least one exon with 100% accuracy.

DOE Grant No. DE-FG03-97ER62383.

Foundations for a Syntactic Pattern-Recognition System for Genomic DNA Sequences: Languages, Automata, Interfaces, and Macromolecules

David B. Searls and G. Christian Overton[1]
SmithKline Beecham Pharmaceuticals; King of Prussia, PA 19406
610/270-4551, Fax: -5580, *searldb@sb.com*
[1]Department of Genetics; University of Pennsylvania; Philadelphia, PA 19104

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 probably 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—bioWidget—for the rapid development and deployment of graphical user interfaces (GUIs) designed for the scientific visualization of molecular, cellular and genomics information. The overarching philosophy behind bioWidgets is componentry: that is, the creation of adaptable, reusable software, deployed in modules that are easily incorporated in a variety of applications and in such a way as to promote interaction between those applications. This is in sharp distinction to the common practice of developing dedicated applications. The bioWidgets project additionally focuses on the development of specific applications based on bioWidget componentry, including chromosomes, maps, and nucleic acid and peptide sequences.

The current set of bioWidgets has been implemented in Java with the goal in mind of delivering local applications and distributed applets via Intranet/Internet environments as required. The immediate focus is on developing interfaces for information stored in distributed heterogeneous databases such as GDB, GSDB, Entry, and ACeDB. The issues we are addressing are database access, reflecting database schemas in bioWidgets, and performance. We are also directing our efforts into creating a consortium of bioWidget developers and end-users. This organization will create standards for and encourage the development of bioWidget components. Primary participants in the consortium include Gerry Rubin (UC Berkeley) and Nat Goodman (Jackson Labs).

DOE Grant No. DE-FG02-92ER61371.

Relevant Publications

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 Mutation 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).

Analysis and Annotation of Nucleic Acid Sequence

David J. States, Ron Cytron, Pankaj Agarwal, and Hugh Chou
Institute for Biomedical Computing; Washington University; St. Louis, MO 63108
314/362-2134, Fax: -0234, *states@ibc.wustl.edu*
http://www.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 parameterized 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.

DOE Grant No. DE-FG02-94ER61910.

Gene Recognition, Modeling, and Homology Search in GRAIL and genQuest

Ying Xu, Manesh Shah, J. Ralph Einstein, Sherri Matis, Xiaojun Guan, Sergey Petrov, Loren Hauser,[1] Richard J. Mural,[1] and **Edward C. Uberbacher**
Computer Science and Mathematics and [1]Biology Divisions; Oak Ridge National Laboratory; Oak Ridge, TN 37831
Uberbacher: 423/574-6134, Fax: -7860, *ube@ornl.gov*
http://compbio.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 e-mail 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 rec-

ognition 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 grailsrv.lsd.ornl.gov (124.167.140.21). Both GRAIL and genQuest are accessible over the World Wide Web (URL http://compbio.ornl.gov). Communications with the GRAIL staff should be addressed to GRAILMAIL@ornl.gov.

DOE Contract No. DE-AC05-840R21400.

Informatics Support for Mapping in Mouse-Human Homology Regions

Edward Uberbacher, Richard Mural,[1] Manesh Shah, Loren Hauser,[1] and Sergey Petrov
Computer Science and Mathematics Division and [1]Biology Division; Oak Ridge National Laboratory; Oak Ridge, TN 37831
423/574-6134, Fax: -7860, *ube@ornl.gov*

The purpose of this project is to develop databases and tools for the Oak Ridge National Laboratory (ORNL) Mouse-Human Mapping Project, including the construction of a mapping database for the project; tools for managing and archiving cDNAs and other probes used in the laboratory; and analysis tools for mapping, interspecific backcross, and other needs. Our initial effort involved installing and developing a relational SYBASE database for tracking samples and probes, experimental results, and analyses. Recent work has focused on a corresponding ACeDB implementation containing mouse mapping data and providing numerous graphical views of this data. The initial relational database was constructed with SYBASE using a schema modeled on one implemented at the Lawrence Livermore National Laboratory (LLNL) center; this was because of documentation available for the LLNL system and the opportunity to maximize compatibility with human chromosome 19 mapping. (Major homologies exist between human chromosome l9 and mouse chromosome 7, the initial focus of the ORNL work.)

With some modification, our ACeDB implementation was modeled somewhat on the Lawrence Berkeley National Laboratory (LBNL) chromosome 21 ACeDB system and designed to contain genetic and physical mouse map data as well as homologous human chromosome data. The usefulness of exchanging map information with LLNL (human chromosome 19) and potentially with other centers has led to the implementation of procedures for data export and the import of human mapping data into ORNL databases.

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 database browser to view human chromosome l9 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 interspecific backcross data and YAC restriction mapping have been implemented.

We would like to acknowledge use of ideas from the LLNL and LBNL Human Genome Centers.

DOE Contract No. DE-AC05-840R21400.

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
510/486-5041, Fax: -4004, *mdzorn@lbl.gov*
http://www-hgc.lbl.gov/submitr.html

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.

DOE Contract No. DE-AC03-76SF00098.

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

Charles C. Carlson
The Exploratorium; San Francisco, CA 94123
415/561-0319, Fax: -0307; *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 support 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.

DOE Grant No. DE-FG03-93ER61583.

Documentary Series for Public Broadcasting

Graham Chedd and **Noel Schwerin**
Chedd-Angier Production Company; Watertown, MA 02172
617/926-8300, Fax: -2710

Designed as a 4-hour documentary series for Public Broadcasting, *Genetics in Society* (working title) will explore the ethical, legal, and social implications of genetic technology. Currently funded and in production for a 90-minute special (*Testing Family Ties*), the first program profiles several individuals and families as they confront genetic tests and the information they generate. One high-risk cancer family struggles to make sense of their genetic legacy as it debates prophylactic surgery and whether or not to test for *BRCA*1 and *BRCA*2. In a family without that family risk, news of the Ashkenazi *BRCA*1 finding pushes an anxious Jewish woman to demand testing for herself and her young daughter. In another, a woman chooses to carry to term her prenatally diagnosed Cystic Fibrosis twins, despite social and personal pressures. In a third, a scientist researching the so-called "obesity gene" at a biotech company debates the proper "marketing" of his research and confronts the larger questions it raises about what should be considered "normal" and what constitutes therapy vs enhancement.

Testing Family Ties will explore not only what genetic technology does—in testing, drug development, and potential therapy—but what it means to our sense of self, family, and future and to our concepts of health and normality.

Depending on outstanding funding requests, *Genetics in Society* will be broadcast in the Fall of 1996 or the Winter of 1997 on PBS. Noel Schwerin is Producer/Director. Graham Chedd is Executive Producer.

DOE Grant No. DE-FG06-95ER61995.

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
913/588-6043, Fax: -4060, *collins@ukanvm.cc.ukans.edu*
http://www.kumc.edu/GEC

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 1500 colleagues and the general public have participated in teacher workshops, and over 56,000 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.

DOE Grant No. DE-FG02-92ER61392.

Human Genome Education Program

Lane Conn
Human Genome Education Program; Stanford Human Genome Center; Palo Alto, CA 94304
415/812-2003, Fax: -1916, *lconn@toolik.stanford.edu*

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.

DOE Grant No. DE-FG03-96ER62161.

Your World/Our World–Biotechnology & You: Special Issue on the Human Genome Project

Jeff Davidson and **Laurence Weinberger**
Pennsylvania Biotechnology Association; State College, PA 16801
814/238-4080, Fax: -4081, *73150.1623@compuserve.com*

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.

DOE Grant No. DE-FG02-95ER62107.

The Human Genome Project and Mental Retardation: An Educational Program

Sharon Davis
Department of Research and Program Services; The Arc of the United States; Arlington, TX 76010
817/261-6003, Fax: /277-3491, *sdavis@metronet.com*
http://The Arc.org/welcome.html

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.

DOE Grant No. DE-FG03-96ER62162.

Pathways to Genetic Screening: Molecular Genetics Meets the High-Risk Family

Troy Duster and Diane Beeson[1]
Institute for the Study of Social Change; University of California; Berkeley, CA 94705
510/642-0813, Fax: /8674, *nitrogn@violet.berkeley.edu*
[1]Department of Sociology; California State University; Hayward, CA 94542

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.

DOE Grant No. DE-FG03-92ER61393.

Intellectual Property Issues in Genomics

Rebecca S. Eisenberg
University of Michigan Law School; Ann Arbor, MI 48109
313/763-1372, Fax: -9375, *rse@umich.edu*

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 communi-

cation 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.

DOE Grant No. DE-FG02-94ER61792.

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.

DOE Grant No. DE-FG02-95ER61974.

A Hispanic Educational Program for Scientific, Ethical, Legal, and Social Aspects of the Human Genome Project

Margaret C. Jefferson and **Mary Ann Sesma**[1]
Department of Biology and Microbiology; California State University; Los Angeles CA 90032
213/343-2059, Fax: -2095, *mjeffer@flytrap.calstatela.edu*
http://vflylab.calstatela.edu/hgp
[1]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.

DOE Grant No. DE-FG03-94ER61797.

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, and Marion Verp
Department of Obstetrics and Gynecology and MacLean Center for Clinical Medical Ethics; University of Chicago; Chicago, IL 60637
312/702-9300, Fax: -0840, *mm46@midway.uchicago.edu*
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 and address issues specific to their area, developing course outlines, bibliography, and methodology based on grand rounds given by national expert.

Primary Care Area

Pediatrics: Genetics expert: Stephen Friend, Ethics Expert: Lainie F. Ross + fellow

Obstetrics/Gynecology: Genetics expert: Joe Leigh Simpson, Ethics Expert: Marion Verp + fellow

Medicine: Genetics expert: Tom Caskey, Ethics Expert: Greg Sachs + fellow

Family medicine: Genetics expert: Noralane Lindor, Ethics Expert: Robert Moss + fellow

Nursing: Genetics expert: Mira Lessick, Ethics Expert: 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.

DOE Grant No. DE-FG02-95ER61990.

References

[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**, 24142 (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**, 1 19-21 (1983).

[6]George, J.B., Genetics: Challenges for nursing education, *J Ped Nursing* **7**, 5-8, (1992).

Nontraditional Inheritance: Genetics and the Nature of Science; Instructional Materials for High School Biology

Joseph D. McInerney and **B. Ellen Friedman**
Biological Sciences Curriculum Study; Colorado Springs, CO 80918
719/531-5550, Fax: -9104, *jmcinerney@cc.colorado.edu*

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.

The Biological Sciences Curriculum Study (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.

DOE Grant No. DE-FG03-95ER61989.

The Human Genome Project: Biology, Computers, and Privacy: Development of Educational Materials for High School Biology

Joseph D. McInerney, **Lynda B. Micikas**, and B. Ellen Friedman
Biological Sciences Curriculum Study; Colorado Springs, CO 80918
719/531-5550, Fax: -9104, *jmcinerney@cc.colorado.edu*

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.

The Biological Sciences Curriculum Study (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.

DOE Grant No. DE-FG03-93ER61584.

Involvement of High School Students in Sequencing the Human Genome

Maureen M. Munn, **Maynard V. Olson**, and **Leroy Hood**
Department of Molecular Biotechnology; University of Washington; Seattle, WA 98195
206/616-4538, Fax: /685-7344, *mmunn@u.washington.edu*

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 biotinlabeled 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 with 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.

DOE Grant No. DE-FG03-96ER62175.

The Gene Letter: A Newsletter on Ethical, Legal, and Social Issues in Genetics for Interested Professionals and Consumers

Philip J. Reilly, **Dorothy C. Wertz**, and **Robin J.R. Blatt**[1]
The Shriver Center for Mental Retardation; Division of Social Science, Ethics and Law; Waltham, MA 02254
617/642-0230, Fax: /893-5340, *preilly@shriver.org*
[1]Also at Massachusetts Department of Public Health, Boston, MA
http://www.shriver.org

We propose to develop a newsletter on ELSI-related issues for dissemination to a broad general audience of professionals and consumers. No such focussed public newsletter currently exists. Entitled *The Gene Letter*, the newsletter will be distributed monthly on-line, through the Internet. Updated weekly on the Internet, it will be poised to react in a timely fashion to new developments in science, law, medicine, ethics, and culture. The newsletter does not propose to provide comprehensive education in genetics for the American public, but rather to begin an information network that interested people can use for further information. It will be the most widely-distributed newsletter on ELSI genetics in the world, with the largest consumer readership. Features will be largely informational and will include new scientific/medical developments and attendant ELSI issues, new court decisions, legislation, and regulations, balanced responses to new concerns in the media, and new developments related to health that may be of interest to health care providers and consumers. Features will present balanced opinions. An editorial board will review each issue, prior to publication, for cultural sensitivity, emphasis, balance, and concerns of persons with disabilities. *The Gene Letter* will also include factual information on upcoming events, new ELSI research, where to find genetics on the Internet, new publications (annotated), and where to find further information about each feature. Readers will be invited to send letters, queries, news, bibliography, comments, and consumer concerns either on *The Gene Letter* Internet chatroom or in hard copy. A hard copy of the first on-line issue will be used to assess readers' needs and interests. It will be distributed to 500 community college students representing blue-collar ethnic groups, and to 2000 members of a broad general audience.

A special evaluation of readers' knowledge and ethical/social concerns raised by *The Gene Letter* will take place at the end of the second year in order to assess outcome. It is our intention that *The Gene Letter* become self-supporting after two years.

DOE Grant No. DE-FG02-96ER62174.

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; KPFA-FM; Berkeley, CA 94704
510/848-6767 ext 235, Fax: /883-0311, *strp@aol.com*

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."

DOE Grant No. DE-FG03-95ER62003.

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; Washington, DC 20005
202/326-6453, Fax: /371-9849, *msosa@aaas.org*

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.

DOE Grant No. DE-FG02-95ER61988.

The Community College Initiative

Sylvia J. Spengler and Laurel Egenberger
Lawrence Berkeley National Laboratory; Berkeley, CA 94720
510/486-4879, Fax: -5717, *sjspengler@lbl.gov*
http://csee.lbl.gov/cup/ccibiotech/Index.html

The Community College Initiative prepares community college students for work in biotechnology. A combined effort of Lawrence Berkeley National Laboratory (LBNL) and the California Community Colleges, we aim to develop mechanisms to encourage students to pursue science studies, to participate in forefront laboratory research, and to gain work experience. The initiative is structured to upgrade the skills of students and their instructors through four components.

Summer Student Workshops: Four weeks summer residential programs for students who have completed the first year of the biotechnology academic program. Ethical, legal

and social concerns are integrated into the laboratory exercises and students learn to identify commonly shared values of the scientific community as well as 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, with emphasis on curriculum development in current topics in ethical, legal, and social issues in science.

Sabbatical Fellowships: For community college instructors to provide investigative and field experience in research laboratories. During the fellowship, teachers also assist in development of student summer research activities.

Summer Faculty-Student Teams: Post-fellowship faculty and biotechnology students who have finished their second year of study team on a research project.

Genome Educators

Sylvia Spengler and **Janice Mann**
Human Genome Program; Life Sciences Division; Lawrence Berkeley National Laboratory; Berkeley, CA 94720
510/486-4879, Fax: -5717, *sjspengler@lbl.gov* or *jlmann@lbl.gov*
http://www.lbl.gov/Education/Genome

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.

Getting the Word Out on the Human Genome Project: A Course for Physicians

Sara L. Tobin and **Ann Boughton**[1]
Department of Biochemistry and Molecular Biology; Center for Biomedical Ethics; Stanford University; Palo Alto, CA 94304-1709
415/725-2663, Fax: -6131, *tobinsl@leland.stanford.edu*
[1]Thumbnail Graphics; 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.

DOE Grant No. DE-FG03-96ER62172.

The Genetics Adjudication Resource Project

Franklin M. Zweig
Einstein Institute for Science, Health, and the Courts; Bethesda, MD 20814
301/961-1949, Fax: /913-0448, *einshac@aol.com*
http://www.ornl.gov/courts

The Einstein Institute for Science, Health, and 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 and others in science and policymaking 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 multipurpose 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.

DOE Grant No. DE-FG02-96ER62081.

Alexander Hollaender Distinguished Postdoctoral Fellowships

Linda Holmes and **Eugene Spejewski**
Oak Ridge Institute for Science and Education; Oak Ridge, TN 37831-0117
423/576-3192, Fax: /241-5220, *holmesl@orau.gov* or *alexpgm@orau.gov*
http://www.orau.gov/oher/hollaend.htm

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 2 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 1999 fellowship cycle is January 15, 1998. 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: /241-5220).

DOE Contract No. DE-AC05-760R00033.

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
Life Sciences Division; Oak Ridge National Laboratory; Oak Ridge, TN 37830
423/576-6669, Fax: /574-9888, *bkq@ornl.gov*
http://www.ornl.gov/hgmis

The Human Genome Management Information System (HGMIS), established in 1989, provides information about the international Human Genome Project in print and World Wide Web formats to both technical and general audiences. HGMIS is sponsored by the Human Genome Program Task Group of the DOE Office of Biological and Environmental Research to help fulfill DOE's commitment to informing scientists, policymakers, and the public about the program's funded research and the context in which the research is conducted. Several HGMIS products, including the Web sites and newsletter, have won technical and electronic communication awards.

HGMIS goals center on facilitating research at the interface of genomics and other biological disciplines that seek revolutionary solutions to biological, environmental, and biomedical challenges. By communicating information about the Human Genome Project and its impact, HGMIS increases the use of project-generated resources, reduces duplicative research efforts, and fosters collaborations and contributions to biology from other research disciplines.

Furthermore, communicating scientific and societal issues to nonscientist audiences contributes to increased science literacy, thus laying a foundation for more informed decision making and public-policy development. For example, since 1995 HGMIS has been participating in a project to educate the judiciary about the basics of genetics and gene testing. The aim is to prepare judges for the flood of cases involving genetic evidence that soon will enter the nation's courtrooms.

Information Resources

In keeping with its goals, HGMIS produces the following information resources in print and on the Web:

Human Genome News (HGN). A quarterly forum for interdisciplinary information exchange, *HGN* uniquely presents a broad spectrum of topics related to the Human Genome Project in a single publication. Articles feature topics that include project goals, progress, and direction; available resources; applications of project data and resources to provide a better understanding of biological processes; related or spinoff programs; medical uses of genome data; ethical, legal, and social considerations; legislative updates; other publications; meeting calendars; and funding information. Most *HGN* articles also contain sources of additional information. In May 1997, DOE acknowledged the newsletter's value by presenting an exceptional service award to *HGN*'s managing editor at a symposium celebrating 50 years of biological and environmental research.

Among 14,000 domestic and foreign *HGN* subscribers are genome and basic researchers at universities, national laboratories, nonprofit organizations, and industrial facilities; educators; industry representatives; legal personnel; ethicists; students; genetic counselors; medical profession-

als; science writers; and other interested individuals. All 41 issues of *HGN*, indexed and searchable, are accessible via the HGMIS Web site.

Other Publications. HGMIS also produces the DOE *Primer on Molecular Genetics*, progress reports on the DOE Human Genome Program, Santa Fe contractor-grantee workshop proceedings, 1-page topical handouts, and other related resource documents. Expanded and revised by HGMIS from an earlier DOE document, the DOE *Primer on Molecular Genetics* continues to be in demand. It is used as a handout for genome centers; a resource for new staff training by companies that make products for genome scientists; and an educational tool for teachers, genetic counselors, and such organizations as high schools, universities, and medical schools for student and continuing-education curricula. More than 35,000 hard copies have been distributed. The primer also is available in several formats at the HGMIS Web site, including an Adobe Acrobat version that can be used to print "originals" from users' printers.

Distribution of Documents. HGMIS has distributed more than 65,000 copies of items requested by subscribers, meeting attendees, and managers of genetics meetings and educational events. These items include HGN, program and workshop reports, DOE-NIH 5-year plans, DOE *Primer on Molecular Genetics*, and *To Know Ourselves*. On request, HGMIS supplies multiple copies of publications for meetings and educational purposes.

Electronic Communications. In November 1994, HGMIS began producing a comprehensive, text-based Web server called Human Genome Project Information, which is devoted to topics relating to the science and societal issues surrounding the genome project. In July 1997, this site was divided to better serve the two diverse audience categories that represent the majority of users: scientists and the public. The sites contain more than 1700 text files that are accessed over 1.2 million times each year. Each month, about 10,000 host computers connect to the HGMIS sites directly and through more than 1000 other Web sites. In addition, HGMIS links to the National Institutes of Health and international Human Genome Organisation sites, as well as to sites dedicated to education and to the ethical, legal, and social implications of the Human Genome Project.

All HGMIS publications are published on the Web site, along with such DOE-sponsored documents as *Your Genes, Your Choices*; the Genetic Privacy Act; and historical and other documents pertaining to the Human Genome Project. HGMIS collaborates with the Einstein Institute for Science, Health, and the Courts to produce *CASOLM*, the online magazine for judicial education in genetics and biomedical issues. HGMIS also maintains the Genetics section of the Virtual Library from CERN (Switzerland) and the DOE Human Genome Program pages and moderates the BioSci Human Genome Newsgroup.

Information Source

HGMIS answers individual questions and supplies general information about the Human Genome Project by telephone, fax, and e-mail and, as appropriate, links scientists with questions to appropriate Human Genome Project contacts. HGMIS staff exchange ideas and suggestions with investigators, industry representatives, and others when attending occasional scientific conferences and genome-related meetings and displaying the DOE Human Genome Project traveling exhibit. HGMIS staff also make presentations on the Human Genome Project to educational, judicial, and other groups.

HGMIS resources serve as a primary source for the popular media and for discipline-specific publications that broaden the distribution of genome project information by extracting and reprinting from HGMIS resources and by linking to various parts of the HGMIS Web site.

HGMIS continuously monitors changes in the direction of the international Human Genome Project and searches for ways to strengthen the content relevancy of the newsletter, the Web site, and other services.

DOE Contract No. DE-AC05-96OR22464.

Human Genome Program Coordination

Sylvia J. Spengler
Lawrence Berkeley National Laboratory; Berkeley CA 94720
510/486-4879, Fax: -5717, *sjspengler@lbl.gov*
http://www.lbl.gov/Education/ELSI

The DOE Human Genome Program of the Office of Health and Environmental Research (OHER) has developed a number of tools for management of the Program. Among these was the Human Genome Coordinating Committee (HGCC), established in 1988. In 1996, the HGCC was expanded to a broader vision of the role of genomic technologies in OHER programs, and the name was changed to reflect this broadening. The HGCC is now the Biotechnology Forum. The Forum is chaired by the Associate Director, OHER. Members of the Human Genome Program Management Task group are ex officio members, as are members of the Health and Environmental Research Advisory Committee's subcommittee on the Human Genome. Responsibilities of the Forum include: assisting OHER in overall coordination of DOE-funded genome research; facilitating the development and dissemination of novel genome technologies; recommending establishment of ad hoc task groups in specific areas, such as informatics,

technologies, model organisms; and evaluation of progress and consideration of long-term goals. Members also serve on the Joint DOE-NIH Subcommittee on the Human genome, for interagency coordination. The coordination group also participates in interface programs with other facilities and provides scientific support for development of other OHER goals, as requested.

Support of Human Genome Program Proposal Reviews

Walter Williams
Education/Training Division; Oak Ridge Institute for Science and Education; Oak Ridge, TN 37831-0117
423/576-4811, Fax: /241-2727, *williamw@orau.gov*

The Oak Ridge Institute for Science and Education (ORISE), operated by Oak Ridge Associated Universities, provides assistance to the DOE Office of Health and Environmental Research in the technical review of proposals submitted in response to solicitations by the DOE Human Genome Program. ORISE staff members create and maintain a database of all proposal information; including abstracts, relevant names and addresses, and budget data. This information is compiled and presented to proposal reviewers. Before review meetings, ORISE staff members make appropriate hotel and meeting arrangements, provide each reviewer with proposal copies and evaluation guidelines, and coordinate reviewer travel and honoraria payment. Onsite meeting support includes collecting all reviewer evaluation forms and scores, entering reviewer scores into the database, preparing appropriate reports, providing onsite computer support, and handling all logistical issues. Other support includes assistance with program advertising and preparation of reviewer comments following each review. ORISE may also assist with pre- and post-review activities related to conferences, seminars, and site visits.

DOE Contract No. DE-AC05-76OR00033.

Former Soviet Union Office of Health and Environmental Research Program

James Wright
Education/Training Division; Oak Ridge Institute for Science and Education; Oak Ridge, TN 37831-0117
423/576-1716, Fax: /241-2727, *wrightj@orau.gov*

The Former Soviet Union Office of Health and Environmental Research Program, sponsored by the U.S. Department of Energy, Office of Health and Environmental Research, recognizes outstanding scientists in the field of health and environmental research from the independent states of the former Soviet Union. The program fosters the international exchange of new ideas and innovative approaches in health and environmental research; strengthens ties and encourages continuing collaboration among Russians and U.S. scientists; and establishes and maintains environmental research capability in the former Soviet Union. The program has supported more than 23 Russian principal investigators and approximately 110 other research associates in Moscow, St. Petersburg, and Novosibirsk. More importantly, the program has enabled many high quality Russian biological, genome informatics, physical mapping and mutagenesis detection, human genetics,, biochemistry, DNA sequencing technology, protein analysis, molecular genetics, and other related research infrastructures to continue operating in an uncertain economic environment.

DOE Contract No. DE-AC05-76OR00033.

1996 Phase I

An Engineered RNA/DNA Polymerase to Increase Speed and Economy of DNA Sequencing

Mark W. Knuth
Promega Corporation; Madison, WI 53711-5399
608/274-4330, Fax: /277-2601

DNA sequence information is the cornerstone for considerable experimental design and analysis in the biological sciences. The proposed studies will focus on advancing DNA sequencing by creating a new enzyme that eliminates the need for an oligonucleotide primer to initiate DNA synthesis at a defined site, and that can use dideoxy nucleotides for chain termination. The new method should reduce the time and cost required to obtain DNA sequences and enhance the speed and cost effectiveness of current DNA sequencing technologies. Phase I studies will focus on purifying mutant T7 RNA polymerases known to incorporate dNTPs into DNA chains, developing protocols for rapid small scale mutant enzyme purification, evaluating the purified mutants for properties relevant to DNA sequencing, developing facile mutagenesis schemes and producing mutant RNA/DNA polymerases with altered promoter recognition. The results from phase I will provide the foundation for Phase II research, which will focus on refining properties of the mutant by: (1) expanding the number of mutations examined using the purification protocols, assays, and mutagenesis screening methods developed in Phase I and (2) examining the effect of each mutation on enzymatic properties important to DNA sequencing applications, and (3) optimizing conditions for sequencing performance. In Phase III, Promega will commercialize the new mutant enzymes through its own extensive distribution network and by collaborating with major instrumentation firms to adapt the technology to automated DNA sequencing systems.

DOE Grant No. DE-FG02-96ER8226.

Directed Multiple DNA Sequencing and Expression Analysis by Hybridization

Gualberto Ruano
BIOS Laboratories, Inc.; New Haven, CT 06511
800/678-9487 or 203/773-1450, Fax: 800/315-7435 or 203/562-9377

The overall goal of this project is to develop molecular resources with direct applications to either DNA sequence analysis or gene expression analysis in multiplexed formats using sequential hybridization of Peptide Nucleic Acid (PNA) oligomer probes. PNA oligomers hybridize more stably and specifically to cognate DNA targets than conventional DNA oligonucleotides. The Phase I project discussed here is concerned with development of PNA probe technology having direct application either to the directed sequencing process or to gene expression profiling. With regard to directed sequencing, we seek improvements in the three multiply repeated steps associated with this process, namely (1) probe assembly, (2) sequencing reactions, and (3) gel electrophoresis. In PNA hybridization sequencing, sequences are generated directly from the template by multiplex DNA sequencing using anchor primers known to have frequent annealing sites. Electrophoresis is performed en masse for each anchor primer reaction, blotted to nylon membranes and individual sequences are selectively exposed by iterative hybridization to specific 8-mer PNA probes derived from sequences statistically over-represented in expressed DNA and obtained from a pre-synthesized library. Additionally, the same PNA library can be used as a source of hybridization probes for querying expression patterns of specific genes in any cell line or tissue. Specific gene expression can be monitored by coupling gene-specific RT-PCR with hybridization when cDNA products are separated by gel electrophoresis and blotted to nylon membranes. Patterns of gene expression are then resolved by hybridization using PNA oligomers. Bands corresponding to specific genes can be deconvoluted using sequence information from RT-PCR primers and PNA probes. Higher throughput expression analysis can be achieved by multiplexed gel electrophoresis, blotting and iterative probing of RT-PCR reactions with individual PNA probes.

DOE Grant No. DE-FG02-96ER8213.

1996 Phase II

A Graphical Ad Hoc Query Interface Capable of Accessing Heterogeneous Public Genome Databases

Joseph Leone
CyberConnect Corporation; Storrs, CT 06268
860/486-2783, Fax: /429-2372

The interoperability of public genome databases is expected to be crucial in making the Human Genome Project a success. This project will develop software tools in which users in the genome community can learn or examine public genome database schemes in a relatively short time and can produce a correct Structured Query Language (SQL) expression easily. In Phase I, a concept system was constructed and the effectiveness of formulating ad hoc queries graphically was demonstrated. Phase II will focus on transforming the concept system into a product that is robust and portable. Two types of computer programs will be developed. One is a client program which is to be distributed to community users who intend to access public genomic databases and link them with local databases. The other is a server program and a suite of software tools designed to be used by those genome centers which intend to make their databases publicly accessible.

DOE Grant No. DE-FG02-95ER81906.

Low-Cost Automated Preparation of Plasmid, Cosmid, and Yeast DNA

Tuyen Nguyen, Randy F. Sivila, Joshua P. Dyer, and
William P. MacConnell
MacConnell Research Corporation; San Diego, CA 92121
619/452-2603, Fax: -6753

MacConnell Research currently manufactures and sells a low cost automated bench-top instrument that can purify up to 24 samples of plasmid DNA simultaneously in one hour at a cost of $0.65 per sample and under $8000 for the instrument. The patented instrument uses a form of agarose gel electrophoresis to purify the plasmid DNA and electroelutes into approximately a 20 +l volume. The instrument has many advantages over other robotic and manual methods including the fact that is it two times faster, at least six times less expensive, much smaller in size, easier to operate, less cost per sample, and results in DNA pure enough for direct use in fluorescent automated sequencing. The instrument process begins with bacterial culture which is loaded directly into a disposable cassette in the machine.

In Phase II work we are developing an instrument which simultaneously purifies plasmid DNA from up to 192 (2 X 96) bacterial samples in 1.5 hours. Prototypes of this instrument thus far constructed have allowed the purification of 3–7 micrograms of high purity plasmid DNA per lane from 1.5 ml of bacterial culture. We have attempted to optimize all of the: instrument electrophoretic run parameters, lysis chemistry, lysis reagent delivery devices, reagent storage at room temperature, desalting processes and overall instrument mechanical and electronic control. Instrument prototypes have also been used to prepare cosmid or yeast DNA in quantities of 1–5 micrograms per cassette lane. Trials thus far have yielded plasmid DNA of sufficient purity for direct use in automated fluorescent and manual sequencing as well as other molecular biology protocols. We have studied the purity of the resulting DNA when directly sequenced on a Licor 4000 Long Reader and ABI 373A automated DNA sequencers. Results from the Licor 4000 instrument give routine read lengths of >850 base pairs with 98% accuracy while ABI 373A reads generally exceed 400 base pairs with similar accuracy.

The proposed 2 X 96-channel instrument will purify up to 1200 plasmid DNA preps per eight hour day. It will significantly reduce the cost and technician labor of high throughput plasmid DNA purification for automated sequencing and mapping.

DOE Grant No. DE-FG03-94ER81802/A000.

GRAIL-GenQuest: A Comprehensive Computational Framework for DNA Sequence Analysis

Ruth Ann Manning
ApoCom, Inc.; Oak Ridge, TN 37830
423/482-2500, Fax: /220-2030

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. The client-server DNA sequence analysis system GRAIL is the most accurate and widely used computer-based system for locating and characterizing genes in DNA sequences, but it is not accessible to many biotechnology environments. The GRAIL client software and graphical displays have been developed for high-end UNIX-based computer workstations. Such workstations are standard equipment in universities and large companies, but personal computers (PCs) and Macintosh computers are the prevalent technology within the biotechnology community. This Phase I project will design Macintosh- and Windows-based client graphical user interface prototypes for GRAIL.

The growth of DNA databases is expected to continue at a fast pace in the attempt to sequence the human genome completely by the year 2005. Parallel processing is a viable solution to handle searching through the ever-increasing volume of data. During Phase I, genQuest—the sequence comparison server portion of the GRAIL system—will be parallelized for shared-memory platforms and will use PVM[1] for the development of genQuest servers on networks of PCs and workstations and other innovative, high-performance computer architectures.

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 will improve the cost-performance ratio and make parallel searches more affordable to the biotechnology community using general multipurpose hardware.

DOE Grant No. DE-FG02-95ER81923.

[1]The Parallel Virtual Machine (PVM) message-passing library allows a collection of UNIX-based computers to function as a single multiple-processor supercomputer.

Projects Completed FY 1994–95

Projects in this section have been completed or did not receive support through the DOE Human Genome Program in FY 1996.

Sequencing

Sequencing by Hybridization: Methods to Generate Large Arrays of Oligonucleotides
Thomas M. Brennan

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

Genomic Instrumentation Development: Detection Systems for Film and High-Speed Gel-Less Methods
Jack B. Davidson and Robert S. Foote

Single-Molecule Detection Using Charge-Coupled Device Array Technology
M. Bonner Denton, Richard Keller, Mark E. Baker, Colin W. Earle, and David A. Radspinner

Coupling Sequencing by Hybridization with Gel Sequencing for Inexpensive Analysis of Genes and Genomes
Radoje Drmanac, Snezana Drmanac, and Ivan Labat

Physical Structure and DNA Sequence of Human Chromosomes
Glen A. Evans

Using Scanning Tunneling Microscopy to Sequence the Human Genome
Thomas L. Ferrell, Robert J. Warmack, David P. Allison, K. Bruce Jacobson, Gilbert M. Brown, and Thomas G. Thundat

DNA Sequence Analysis by Solid-Phase Hybridization
Robert S. Foote, Richard A. Sachleben, and K. Bruce Jacobson

DNA Sequencing Using Stable Isotopes
K. Bruce Jacobson, Heinrich F. Arlinghaus, Gilbert M. Brown, Robert S. Foote, Frank W. Larimer, Richard A. Sachleben, Norbert Thonnard, and Richard P. Woychik

Preparation of Oligonucleotide Arrays for Hybridization Studies
Michael C. Pirrung, Steven W. Shuey, David C. Lever, Lara Fallon, J.-C. Bradley, and William P. Hawe

Improvement and Automation of Ligation-Mediated Genomic Sequencing
Arthur D. Riggs and Gerd P. Pfeifer

*Analysis of a 53-Kb Nucleotide Sequence from the Right Genome Terminus of the Variola Major Virus Strain India-1967
Sergei N. Shchelkunov, Vladimir M. Blinov, Sergei M. Resenchuk, Alexei V. Totmenin, Viktor N. Krasnykh, Ludmilla V. Olenina, **Oleg I. Serpinsky**, and Lev S. Sandakhchiev

A High-Speed Automated DNA Sequencer
Lloyd M. Smith

Characterization and Modification of DNA Polymerases for Use in DNA Sequencing
Stanley Tabor

Mapping

*Toward Cloning Human Chromosome 19 in Yeast Artificial Chromosomes
Inga P. Arman, Alexander B. Devin, Svetlana P. Legchilina, Irina G. Efimenko, Marina E. Smirnova, and Dina V. Glazkova

A Panel of Mouse-Human Monochromosomal Hybrid Cell Lines, Each Containing a Single Different Tagged Human Chromosome
Arbansjit K. Sandhu, G. Pal Kaur, and **Raghbir S. Athwal**

*Preparation of a Set of Molecular Markers for Human Chromosome 5 Using G+C–Rich and Functional Site-Specific Oligonucleotides
M.L. Filipenko, A.I. Muravlev, E.I. Jantsen, V.V. Smirnova, N.A. Chikaev, V.P. Mishin, and M.A. Ivanovich

An Improved Method for Producing Radiation Hybrids Applied to Human Chromosome 19
Cynthia L. Jackson and Hon Fong L. Mark

Completed Projects

Construction of a Human Genome Library Composed of Multimegabase Acentric Chromosome Fragments
Michael J. Lane, **Peter Hahn**, and **John Hozier**

Reagents for Understanding and Sequencing the Human Genome
J.R. Korenberg, X-N. Chen, S. Mitchell, S. Gerwehr, Z. Sun, D. Noya, R. Hubert, U-J. Kim, H. Shizuya, X. Wu, J. Silva, B. Birren, T.J. Hudson, P. de Jong, E. Lander, and M. Simon

Development of Diallelic Marker Maps Using PCR/OLA
Deborah A. Nickerson and **Pui-Yan Kwok**

Multiplex Mapping of Human cDNAs
William C. Nierman, Donna R. Maglott, and Scott Durkin

Physical Mapping in Preparation for DNA Sequencing
Andreas Gnirke, Regina Lim, Gane Wong, Jun Yu, Roger Bumgarner, and **Maynard Olson**

Construction of a Genetic Map Across Chromosome 21
Elaine A. Ostrander

Integrated Physical Mapping of Human cDNAs
Mihael H. Polymeropoulos

Sequence-Tagged Sites for Human Chromosome 19 cDNAs
Michael J. Siciliano and Anthony V. Carrano

cDNA/STS Map of the Human Genome: Methods Development and Applications Using Brain cDNAs
James M. Sikela, Akbar S. Khan, Arto K. Orpana, Andrea S. Wilcox, Janet A. Hopkins, and Tamara J. Stevens

Physical Structure of Human Chromosome 21
Cassandra L. Smith, Denan Wang, Kaoru Yoshida, Jesus Sainz, Carita Fockler, and Meire Bremer

Physical Mapping of Human Chromosome 16
David F. Callen, Sinoula Apostolou, Elizabeth Baker, Helen Kozman, Sharon A. Lane, Julie Nancarrow, Hilary A. Phillips, Scott A. Whitmore, Norman A. Doggett, John C. Mulley, Robert I. Richards, and **Grant R. Sutherland**

Chromosome Mapping by FISH to Interphase Nuclei
Barbara J. Trask

Flow Karyotyping and Flow Instrumentation Development
Ger van den Engh and **Barbara Trask**

Isolation of Specific Human Telomeric Clones by Homologous Recombination and YAC Rescue
Geoffrey Wahl and Linnea Brody

Informatics

*A Method for Direct Sequencing of Diploid Genomes on Oligonucleotide Arrays: Theoretical Analysis and Computer Modeling
Alexander B. Chetverin

Sampling-Based Methods for the Estimation of DNA Sequence Accuracy
Gary Churchill and **Betty Lazareva**

Computer-Aided Genome Map Assembly with SIGMA (System for Integrated Genome Map Assembly)
Michael J. Cinkosky, Michael A. Bridgers, William M. Barber, Mohamad Ijadi, and James W. Fickett

Informatics for the Sequencing by Hybridization Project
Aleksandar Milosavljevic and **Radomir Crkvenjakov**

Sequencing by Hybridization Algorithms and Computational Tools
Radoje Drmanac, Ivan Labat, and Nick Stavropoulos

HGIR: Information Management for a Growing Map
James W. Fickett, Michael J. Cinkosky, Michael A. Bridgers, Henry T. Brown, Christian Burks, Philip E. Hempfner, Tran N. Lai, Debra Nelson, Robert M. Pecherer, Doug Sorenson, Peichen H. Sgro, Robert D. Sutherland, Charles D. Troup, and Bonnie C. Yantis

Identification of Genes in Anonymous DNA Sequences
Christopher A. Fields and **Carol A. Soderlund**

Algorithms in Support of the Human Genome Project
Dan Gusfield, Jim Knight, Kevin Murphy, Paul Stelling, Lushen Wang, Archie Cobbs, Paul Horton, Richard Karp, and Gene Lawler

BISP: VLSI Solutions to Sequence-Comparison Problems
Tim Hunkapiller, Leroy Hood, Ed Chen, and Michael Waterman

Physical Mapping of DNA Molecules
Richard M. Karp

BIOSCI Electronic Newsgroup Network for the Biological Sciences
David Kristofferson

Multiple Alignment and Homolog Sequence Database Compilation
Hwa A. Lim

Applying Machine Learning Techniques to DNA Sequence Analysis
Jude W. Shavlik, Michiel O. Noordewier, Geoffrey Towell, Mark Craven, Andrew Whitsitt, Kevin Cherkauer, and Lorien Pratt

New Approaches to Recognizing Functional Domains in Biological Sequences
Gary D. Stormo

ELSI

Protecting Genetic Privacy by Regulating the Collection, Analysis, Use, and Storage of DNA and Information Obtained from DNA Analysis
George J. Annas, Leonard H. Glantz, and Patricia A. Roche

"The Secret of Life"
Paula Apsell and **Graham Chedd**

Genome Technology and Its Implications: A Hands-On Workshop for Educators
Diane Baker and **Paula Gregory**

Predicting Future Disease: Issues in the Development, Application, and Use of Tests for Genetic Disorders
Ruth E. Bulger and Jane E. Fullarton

HUGO International Yearbook: Genetics, Ethics, Law, and Society (GELS)
Alex Capron and **Bartha Knoppers**

The Human Genome: Science and the Social Consequences; Interactive Exhibits and Programs on Genetics and the Human Genome
Charles C. Carlson

International Conference Working Group: The Social Costs and Medical Benefits of Human Genetic Information
Betsy Fader

"Medicine at the Crossroads"
George Page and **Stefan Moore**

Pilot Senior Research Fellowship Program: Bioethical Issues in Molecular Genetics
Declan Murphy and **Claudette Cyr Friedman**

Studies of Genetic Discrimination
Marvin Natowicz

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

Mechanical Interactive Exhibits on Biotechnology
Elizabeth Sharpe

Impact of Technology Derived from the Human Genome Project on Genetic Testing, Screening, and Counseling: Cultural, Ethical, and Legal Issues
Ralph W. Trottier, Lee A. Crandall, David Phoenix, Mwalimu Imara, and Ray E. Mosley

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

Completed Projects

Human Genetics and Genome Analysis: A Practical Workshop for Public Policymakers and Opinion Leaders
Jan Witkowski, David A. Micklos, and Margaret Henderson

SBIR Phase I

A Graphical Ad Hoc Query Interface Capable of Accessing Heterogenous Public Genome Databases
J. Clarke Anderson

Techniques for Screening Large-Insert Libraries
Saika Aytay

Interactive DNA Sequence Processing for a Microcomputer
Wayne Dettloff and **Holt Anderson**

High-Performance Searching and Pattern Recognition for Human Genome Databases
Douglas J. Eadline

Estimating, Encoding, and Using Uncertainties in Sequence Data
John R. Hartman

Low-Cost Massively Parallel Neurocomputing for Pattern Recognition in Macromolecular Sequences
John R. Hartman

Electrophoretic Separation of DNA Fragments in Ultrathin Planar-Format Linear Polyacrylamide
Michael T. MacDonell and **Darlene B. Roszak**

An Acoustic Plate Mode DNA Biosensor
Douglas J. McAllister

Piezoelectric Biosensor Using Peptide Nucleic Acids for Triplex Capture
Douglas McAllister

Pedigree Software for the Presentation of Human Genome Information for Genetic Education and Counseling
Charles L. Manske

A High-Spatial-Resolution Spectrograph for DNA Sequencing
Cathy D. Newman

Nonradioactive Detection Systems Based on Enzyme-Fragment Complementation
Peter Richterich

Separation Media for DNA Sequencing
David S. Soane and Herbert H. Hooper

SBIR Phase II

Increased Speed in DNA Sequencing by Utilizing LARIS and SIRIS to Localize Multiple Stable Isotope-Labeled Fragments
Heinrich F. Arlinghaus

Rapid, High-Throughput DNA Sequencing Using Confocal Fluorescence Imaging of Capillary Arrays
David L. Barker and **Jay Flatley**

Spatially Defined Oligonucleotide Arrays
Stephen P. A. Fodor

Site-Specific Endonucleases for Human Genome Mapping
George Golumbeski, Kimberly Knoche, Susanne Selman, im Hartnett, Lydia Hung, and Peter Bayne

High-Performance DNA and Protein Sequence Analysis on a Low-Cost Parallel-Processor Array
John R. Hartman and David L. Solomon

Chemiluminescent Multiprimed DNA Sequencing
Chris S. Martin, Corinne E. M. Olesen, and **Irena Bronstein**

Appendix
Narratives from Large, Multidisciplinary Research Projects

Part 1 of this report contains narratives that represent DOE Human Genome Program research in large, multidisciplinary projects. As a convenience to the reader, these narratives are reprinted without graphics in this appendix. Only the contact persons for these organizations are listed in the Index to Principal and Coinvestigators. To obtain more information on research carried out in these projects, see their contact information or visit the Web sites listed with the narratives.

Joint Genome Institute
Genome Center Sequencing Efforts Merge

Lawrence Livermore National Laboratory
7000 East Avenue, L-452
Livermore, CA 94551

Elbert Branscomb, JGI Scientific Director
510/422-5681
elbert@alu.llnl.gov* or *elbert@shotgun.llnl.gov

http://www.jgi.doe.gov

In a major restructuring of its Human Genome Program, on October 23, 1996, the DOE Office of Biological and Environmental Research established the Joint Genome Institute (JGI) to integrate work based at its three major human genome centers.

The JGI merger represents a shift toward large-scale sequencing via intensified collaborations for more effective use of the unique expertise and resources at Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), and Los Alamos National Laboratory. Elbert Branscomb (LLNL) serves as JGI's Scientific Director. Capital equipment has been ordered, and operational support of about $30 million is projected for the 1998 fiscal year.

With easy access to both LBNL and LLNL, a building in Walnut Creek, California, is being modified. Here, starting in late FY 1998, production DNA sequencing will be carried out for JGI. Until that time, large-scale sequencing will continue at LANL, LBNL, and LLNL. Expectations are that within 3 to 4 years the Production Sequencing Facility will house some 200 researchers and technicians working on high-throughput DNA sequencing using state-of-the-art robotics.

Initial plans are to target gene-rich regions of around 1 to 10 megabases for sequencing. Considerations include gene density, gene families (especially clustered families), correlations to model organism results, technical capabilities, and relevance to the DOE mission (e.g., DNA repair, cancer susceptibility, and impact of genotoxins). The JGI program is subject to regular peer review.

Sequence data will be posted daily on the Web; as the information progresses to finished quality, it will be submitted to public databases.

As JGI and other investigators involved in the Human Genome Project are beginning to reveal the DNA sequence of the 3 billion base pairs in a reference human genome, the data already are becoming valuable reagents for explorations of DNA sequence function in the body, sometimes called "functional genomics." Although large-scale sequencing is JGI's major focus, another important goal will be to enrich the sequence data with information about its biological function. One measure of JGI's progress will be its success at working with other DOE laboratories, genome centers, and non-DOE academic and industrial collaborators. In this way, JGI's evolving capabilities can both serve and benefit from the widest array of partners.

Production DNA Sequencing Begun Worldwide

The year 1996 marked a transition to the final and most challenging phase of the U.S. Human Genome Project, as pilot programs aimed at refining large-scale sequencing strategies and resources were funded by DOE and NIH (see Research Highlights, DNA Sequencing, p. 14). Internationally, large-scale human genome sequencing was kicked off in late 1995 when The Wellcome Trust announced a 7-year, $75-million grant to the private Sanger Centre to scale up its sequencing capabilities. French investigators also have announced intentions to begin production sequencing.

Funding agencies worldwide agree that rapid and free release of data is critical. Other issues include sequence accuracy, types of annotation that will be most useful to biologists, and how to sustain the reference sequence.

The international Human Genome Organisation maintains a Web page to provide information on current and future sequencing projects and links to sites of participating groups (http://hugo.gdb.org). The site also links to reports and resources developed at the February 1996 and 1997 Bermuda meetings on large-scale human genome sequencing, which were sponsored by The Wellcome Trust.

Lawrence Livermore National Laboratory Human Genome Center

Human Genome Center
Lawrence Livermore National Laboratory
Biology and Biotechnology Research Program
7000 East Avenue, L-452
Livermore, CA 94551

Anthony V. Carrano, Director
510/422-5698, Fax: /423-3110, *carrano1@llnl.gov*

Linda Ashworth, Assistant to Center Director
510/422-5665, Fax: -2282, *ashworth1@llnl.gov*

http://www-bio.llnl.gov/bbrp/genome/genome.html

The Human Genome Center at Lawrence Livermore National Laboratory (LLNL) was established by DOE in 1991. The center operates as a multidisciplinary team whose broad goal is understanding human genetic material. It brings together chemists, biologists, molecular biologists, physicists, mathematicians, computer scientists, and engineers in an interactive research environment focused on mapping, DNA sequencing, and characterizing the human genome.

Goals and Priorities

In the past 2 years, the center's goals have undergone an exciting evolution. This change is the result of several factors, both intrinsic and extrinsic to the Human Genome Project. They include: (1) successful completion of the center's first-phase goal, namely a high-resolution, sequence-ready map of human chromosome 19; (2) advances in DNA sequencing that allow accelerated scaleup of this operation; and (3) development of a strategic plan for LLNL's Biology and Biotechnology Research Program that will integrate the center's resources and strengths in genomics with programs in structural biology, individual susceptibility, medical biotechnology, and microbial biotechnology.

The primary goal of LLNL's Human Genome Center is to characterize the mammalian genome at optimal resolution and to provide information and material resources to other in-house or collaborative projects that allow exploitation of genomic biology in a synergistic manner. DNA sequence information provides the biological driver for the center's priorities:

- Generation of highly accurate sequence for chromosome 19.
- Generation of highly accurate sequence for genomic regions of high biological interest to the mission of the DOE Office of Biological and Environmental Research (e.g., genes involved in DNA repair, replication, recombination, xenobiotic metabolism, and cell-cycle control).
- Isolation and sequence of the full insert of cDNA clones associated with genomic regions being sequenced.
- Sequence of selected corresponding regions of the mouse genome in parallel with the human.
- Annotation and position of the sequenced clones with physical landmarks such as linkage markers and sequence tagged sites (STSs).
- Generation of mapped chromosome 19 and other genomic clones [cosmids, bacterial artificial chromosomes (BACs), and P1 artificial chromosomes (PACs)] for collaborating groups.
- Sharing of technology with other groups to minimize duplication of effort.
- Support of downstream biology projects, for example, structural biology, functional studies, human variation, transgenics, medical biotechnology, and microbial biotechnology with know-how, technology, and material resources.

Center Organization and Activities

Completion and publication of the metric physical map of human chromosome 19 in 1995 has led to consolidation of many functions associated with physical mapping, with increased emphasis on DNA sequencing. The center is organized into five broad areas of research and support: sequencing, resources, functional genomics, informatics and analytical genomics, and instrumentation. Each area consists of multiple projects, and extensive interaction occurs both within and among projects.

Sequencing

The sequencing group is divided into several subprojects. The core team is responsible for the construction of sequence libraries, sequencing reactions, and data collection for all templates in the random phase of sequencing. The finishing team works with data produced by the core team to produce highly redundant, highly accurate "finish" sequence on targets of interest. Finally, a team of researchers focuses specifically on development, testing, and implementation of new protocols for the entire group, with an emphasis on improving the efficiency and cost basis of the sequencing operation.

Resources

The resources group provides mapped clonal resources to the sequencing teams. This group performs physical mapping as needed for the DNA sequencing group by using fingerprinting, restriction mapping, fluorescence in situ hybridization, and other techniques. A small mapping effort is under way to identify, isolate, and characterize BAC clones (from anywhere in the human genome) that relate to susceptibility genes, for example, DNA repair. These clones will be characterized and provided for sequencing and at the same time contribute to understanding the biology of the chromosome, the genome, and susceptibility factors. The mapping team also collaborates with others using the chromosome 19 map as a resource for gene hunting.

Functional Genomics

The functional genomics team is responsible for assembling and characterizing clones for the Integrated Molecular Analysis of Gene Expression (called IMAGE) Consortium and cDNA sequencing, as well as for work on gene expression and comparative mouse genomics. The effort emphasizes genes involved in DNA repair and links strongly to LLNL's gene-expression and structural biology efforts. In addition, this team is working closely with Oak Ridge National Laboratory (ORNL) to develop a comparative map and the sequence data for mouse regions syntenic to human chromosome 19.

Informatics and Analytical Genomics

The informatics and analytical genomics group provides computer science support to biologists. The sequencing informatics team works directly with the DNA sequencing group to facilitate and automate sample handing, data acquisition and storage, and DNA sequence analysis and annotation. The analytical genomics team provides statistical and advanced algorithmic expertise. Tasks include development of model-based methods for data capture, signal processing, and feature extraction for DNA sequence and fingerprinting data and analysis of the effectiveness of newly proposed methods for sequencing and mapping.

Instrumentation

The instrumentation group also has multiple components. Group members provide expertise in instrumentation and automation in high-throughput electrophoresis, preparation of high-density replicate DNA and colony filters, fluorescence labeling technologies, and automated sample handling for DNA sequencing. To facilitate seamless integration of new technologies into production use, this group is coupled tightly to the biologist user groups and the informatics group.

Collaborations

The center interacts extensively with other efforts within the LLNL Biology and Biotechnology Research Program and with other programs at LLNL, the academic community, other research institutes, and industry. More than 250 collaborations range from simple probe and clone sharing to detailed gene family studies. The following list reflects some major collaborations.

- Integration of the genetic map of human chromosome 19 with corresponding mouse chromosomes (ORNL).
- Miniaturized polymerase chain reaction instrumentation (LLNL).
- Sequencing of IMAGE Consortium cDNA clones (Washington University, St. Louis).
- Mapping and sequencing of a gene associated with Finnish congenital nephrotic syndrome (University of Oulu, Finland).

Accomplishments

The LLNL Human Genome Center has excelled in several areas, including comparative genomic sequencing of DNA repair genes in human and rodent species, construction of a metric physical map of human chromosome 19, and development and application of new biochemical and mathematical approaches for constructing ordered clone maps. These and other major accomplishments are highlighted below.

- Completion of highly accurate sequencing totaling 1.6 million bases of DNA, including regions spanning human DNA repair genes, the candidate region for a congenital kidney disease gene, and other regions of biological interest on chromosome 19.
- Completion of comparative sequence analysis of 107,500 bases of genomic DNA encompassing the human DNA repair gene *ERCC*2 and the corresponding regions in mouse and hamster. In addition to *ERCC*2, analysis revealed the presence of two previously undescribed genes in all three species. One of these genes is a new member of the kinesin motor protein family. These proteins play a wide variety of roles in the cell, including movement of chromosomes before cell division.
- Complete sequencing of human genomic regions containing two additional DNA repair genes. One of these, *XRCC*3, maps to human chromosome 14 and encodes a protein that may be required for chromosome stability. Analysis of the genomic sequence identified another kinesin motor protein gene physi-

cally linked to *XRCC*3. The second human repair gene, *HHR*23A, maps to 19p13.2. Sequence analysis of 110,000 bases containing *HHR*23A identified six other genes, five of which are new genes with similarity to proteins from mouse, human, yeast, and *Caenorhabditis elegans*.

- Complete sequencing of full-length cDNAs for three new DNA repair genes (*XRCC*2, *XRCC*3, and *XRCC*9) in collaboration with the LLNL DNA repair group.
- Generation of a metric physical map of chromosome 19 spanning at least 95% of the chromosome. This unique map incorporates a metric scale to estimate the distance between genes or other markers of interest to the genetics community.
- Assembly of nearly 45 million bases of *Eco*R I restriction-mapped cosmid contigs for human chromosome 19 using a combination of fingerprinting and cosmid walking. Small gaps in cosmid continuity have been spanned by BAC, PAC, and P1 clones, which are then integrated into the restriction maps. The high depth of coverage of these maps (average redundancy, 4.3-fold) permits selection of a minimum overlapping set of clones for DNA sequencing.
- Placement of more than 400 genes, genetic markers, and other loci on the chromosome 19 cosmid map. Also, 165 new STSs associated with premapped cosmid contigs were generated and added to the physical map.
- Collaborations to identify the gene (*COMP*) responsible for two allelic genetic diseases, pseudoachondroplasia and multiple epiphyseal dysplasia, and the identification of specific mutations causing each condition.
- Through sequence analysis of the 2A subfamily of the human cytochrome P450 enzymes, identification of a new variant that exists in 10% to 20% of individuals and results in reduced ability to metabolize nicotine and the antiblood-clotting drug Coumadin.
- Location of a zinc finger gene that encodes a transcription factor regulating blood-cell development adjacent to telomere repeat sequences, possibly the gene nearest one end of chromosome 19.
- Completion of the genomic and cDNA sequence of the gene for the human Rieske Fe-S protein involved in mitochondrial respiration.
- Expansion of the mouse-human comparative genomics collaboration with ORNL to include study of new groups of clustered transcription factors found on human chromosome 19q and as syntenic homologs on mouse chromosome 7.
- Numerous collaborations (in particular, with Washington University and Merck) continuing to expand the LLNL-based IMAGE Consortium, an effort to characterize the transcribed human genome. The IMAGE clone collection is now the largest public collection of sequenced cDNA clones, with more than 500,000 arrayed clones, 500,000 sequences in public databases, and 10,000 mapped cDNAs.
- Development and deployment of a comprehensive system to handle sample tracking needs of production DNA sequencing. The system combines databases and graphical interfaces running on both Mac and Sun platforms and scales easily to handle large-scale production sequencing.
- Expansion of the LLNL genome center's World Wide Web site to include tables that link to each gene being sequenced, to the quality scores and assembled bases collected each night during the sequencing process, and to the submitted GenBank sequence when a clone is completed. [*http://bbrp.llnl.gov/test-bin/projqcsummary*]
- Implementation of a new database to support sequencing and mapping work on multiple chromosomes and species. Web-based automated tools were developed to facilitate construction of this database, the loading of over 100 million bytes of chromosome 19 data from the existing LLNL database, and automated generation of Web-based input interfaces.
- Significant enhancement of the LLNL Genome Graphical Database Browser software to display and link information obtained at a subcosmid resolution from both restriction map hybridization and sequence feature data. Features, such as genes linked to diseases, allow tracking to fragments as small as 500 base pairs of DNA.
- Development of advanced microfabrication technologies to produce electrophoresis microchannels in large glass substrates for use in DNA sequencing.
- Installation of a new filter-spotting robot that routinely produces $6 \times 6 \times 384$ filters. A $16 \times 16 \times 384$ pattern has been achieved.
- Upgrade of the Lawrence Berkeley National Laboratory colony picker using a second computer so that imaging and picking can occur simultaneously.

Future Plans

Genomic sequencing currently is the dominant function of Livermore's Human Genome Center. The physical mapping effort will ensure an ample supply of sequence-ready clones. For sequencing targets on chromosome 19, this

includes ensuring that the most stable clones (cosmids, BACs, and PACs) are available for sequencing and that regions with such known physical landmarks as STSs and expressed sequenced tags (ESTs) are annotated to facilitate sequence assembly and analysis. The following targets are emphasized for DNA sequencing:

- Regions of high gene density, including regions containing gene families.
- Chromosome 19, of which at least 42 million bases are sequence ready.
- Selected BAC and PAC clones representing regions of about 0.2 million to 1 million bases throughout the human genome; clones would be selected based on such high-priority biological targets as genes involved in DNA repair, replication, recombination, xenobiotic metabolism, cell-cycle checkpoints, or other specific targets of interest.
- Selected BAC and PAC clones from mouse regions syntenic with the genes indicated above.
- Full-insert cDNAs corresponding to the genomic DNA being sequenced.

The informatics team is continuing to deploy broader-based supporting databases for both mapping and sequencing. Where appropriate, Web- and Java-based tools are being developed to enable biologists to interact with data. Recent reorganization within this group enables better direct support to the sequencing group, including evaluating and interfacing sequence-assembly algorithms and analysis tools, data and process tracking, and other informatics functions that will streamline the sequencing process.

The instrumentation effort has three major thrusts: (1) continued development or implementation of laboratory automation to support high-throughput sequencing; (2) development of the next-generation DNA sequencer; and (3) development of robotics to support high-density BAC clone screening. The last two goals warrant further explanation.

The new DNA sequencer being developed under a grant from the National Institutes of Health, with minor support through the DOE genome center, is designed to run 384 lanes simultaneously with a low-viscosity sieving medium. The entire system would be loaded automatically, run, and set up for the next run at 3-hour intervals. If successful, it should provide a 20- to 40-fold increase in throughput over existing machines.

An LLNL-designed high-precision spotting robot, which should allow a density of 98,304 spots in 96 cm^2, is now operating. The goal of this effort is to create high-density filters representing a 10× BAC coverage of both human and mouse genomes (30,000 clones = 1× coverage). Thus each filter would provide ~3× coverage, and eight such filters would provide the desired coverage for both genomes. The filters would be hybridized with amplicons from individual or region-specific cDNAs and ESTs; given the density of the BAC libraries, clones that hybridize should represent a binned set of BACs for a region of interest. These BACs could be the initial substrate for a BAC sequencing strategy. Performing hybridizations in parallel in mouse and human DNA facilitates the development of the mouse map (with ORNL involvement), and sequencing BACs from both species identifies evolutionarily conserved and, perhaps, regulatory regions.

Information generated by sequencing human and mouse DNA in parallel is expected to expand LLNL efforts in functional genomics. Comparative sequence data will be used to develop a high-resolution synteny map of conserved mouse-human domains and incorporate automated northern expression analysis of newly identified genes. Long range, the center hopes to take advantage of a variety of forms of expression analysis, including site-directed mutation analysis in the mouse.

Summary

The Livermore Human Genome Center has undergone a dramatic shift in emphasis toward commitment to large-scale, high-accuracy sequencing of chromosome 19, other chromosomes, and targeted genomic regions in the human and mouse. The center also is committed to exploiting sequence information for functional genomics studies and for other programs, both in house and collaboratively.

Los Alamos National Laboratory Center for Human Genome Studies

Center for Human Genome Studies
Los Alamos National Laboratory
P.O. Box 1663
Los Alamos, NM 87545

Robert K. Moyzis, Director, 1989–97*

*Now at University of California, Irvine

Larry L. Deaven, Acting Director
505/667-3912, Fax: -2891
ldeaven@telomere.lanl.gov

Lynn Clark, Technical Coordinator
505/667-9376, Fax: -2891
clark@telomere.lanl.gov

http://www-ls.lanl.gov/masterhgp.html

Biological research was initiated at Los Alamos National Laboratory (LANL) in the 1940s, when the laboratory began to investigate the physiological and genetic consequences of radiation exposure. Eventual establishment of the national genetic sequence databank called GenBank, the National Flow Cytometry Resource, numerous related individual research projects, and fulfillment of a key role in the National Laboratory Gene Library Project all contributed to LANL's selection as the site for the Center for Human Genome Studies in 1988.

Center Organization and Activities

The LANL genome center is organized into four broad areas of research and support: Physical Mapping, DNA Sequencing, Technology Development, and Biological Interfaces. Each area consists of a variety of projects, and work is distributed among five LANL Divisions (Life Sciences; Theoretical; Computing, Information, and Communications; Chemical Science and Technology; and Engineering Sciences and Applications). Extensive interdisciplinary interactions are encouraged.

Physical Mapping

The construction of chromosome- and region-specific cosmid, bacterial artificial chromosome (BAC), and yeast artificial chromosome (YAC) recombinant DNA libraries is a primary focus of physical mapping activities at LANL. Specific work includes the construction of high-resolution maps of human chromosomes 5 and 16 and associated informatics and gene discovery tasks.

Accomplishments

- Completion of an integrated physical map of human chromosome 16 consisting of both a low-resolution YAC contig map and a high-resolution cosmid contig map. With sequence tagged site (STS) markers provided on average every 125,000 bases, the YAC-STS map provides almost-complete coverage of the chromosome's euchromatic arms. All available loci continue to be incorporated into the map.
- Construction of a low-resolution STS map of human chromosome 5 consisting of 517 STS markers regionally assigned by somatic-cell hybrid approaches. Around 95% mega-YAC–STS coverage (50 million bases) of 5p has been achieved. Additionally, about 40 million bases of 5q mega-YAC–STS coverage have been obtained collaboratively.
- Refinement of BAC cloning procedures for future production of chromosome-specific libraries. Successful partial digestion and cloning of microgram quantities of chromosomal DNA embedded in agarose plugs. Efforts continue to increase the average insert size to about 100,000 bases.

DNA Sequencing

DNA sequencing at the LANL center focuses on low-pass sample sequencing (SASE) of large genomic regions. SASE data is deposited in publicly available databases to allow for wide distribution. Finished sequencing is prioritized from initial SASE analysis and pursued by parallel primer walking. Informatics development includes data tracking, gene-discovery integration with the Sequence Comparison ANalysis (SCAN) program, and functional genomics interaction.

Accomplishments

- SASE sequencing of 1.5 million bases from the p13 region of human chromosome 16.
- Discovery of more than 100 genes in SASE sequences.
- Generation of finished sequence for a 240,000-base telomeric region of human chromosome 7q. From initial sequences generated by SASE, oligonucleotides were synthesized and used for primer walking directly from cosmids comprising the contig map. Complete sequencing was performed to determine what genes, if any, are near the 7q terminus. This intriguing region lacks significant blocks of subtelomeric repeat DNA typically present near eukaryotic telomeres.

- Complete single-pass sequencing of 2018 exon clones generated from LANL's flow-sorted human chromosome 16 cosmid library. About 950 discrete sequences were identified by sequence analysis. Nearly 800 appear to represent expressed sequences from chromosome 16.
- Development of Sequence Viewer to display ABI sequences with trace data on any computer having an Internet connection and a Netscape World Wide Web browser.
- Sequencing and analysis of a novel pericentromeric duplication of a gene-rich cluster between 16p11.1 and Xq28 (in collaboration with Baylor College of Medicine).

Technology Development

Technology development encompasses a variety of activities, both short and long term, including novel vectors for library construction and physical mapping; automation and robotics tools for physical mapping and sequencing; novel approaches to DNA sequencing involving single-molecule detection; and novel approaches to informatics tools for gene identification.

Accomplishments

- Development of SCAN program for large-scale sequence analysis and annotation, including a translator converting SCAN data to GIO format for submission to Genome Sequence DataBase.
- Application of flow-cytometric approach to DNA sizing of P1 artificial chromosome (PAC) clones. Less than one picogram of linear or supercoiled DNA is analyzed in under 3 minutes. Sizing range has been extended down to 287 base pairs. Efforts continue to extend the upper limit beyond 167,000 bases.
- Characterization of the detection of single, fluorescently tagged nucleotides cleaved from multiple DNA fragments suspended in the flow stream of a flow cytometer. The cleavage rate for Exo III at 37°C was measured to be about 5 base pairs per second per M13 DNA fragment. To achieve a single-color sequencing demonstration, either the background burst rate (currently about 5 bursts per second) must be reduced or the exonuclease cleavage rate must be increased significantly. Techniques to achieve both are being explored.
- Construction of a simple and compact apparatus, based on a diode-pumped Nd:YAG laser, for routine DNA fragment sizing.
- Development of a new approach to detect coding sequences in DNA. This complete spectral analysis of coding and noncoding sequences is as sensitive in its first implementations as the best existing techniques.
- Use of phylogenetic relationships to generate new profiles of amino acid usage in conserved domains. The profiles are particularly useful for classification of distantly related sequences.

Biological Interfaces

The Biological Interfaces effort targets genes and chromosome regions associated with DNA damage and repair, mitotic stability, and chromosome structure and function as primary subjects for physical mapping and sequencing. Specific disease-associated genes on human chromosome 5 (e.g., Cri-du-Chat syndrome) and on 16 (e.g., Batten's disease and Fanconi anemia) are the subjects of collaborative biological projects.

Accomplishments

- Identification of two human 7q exons having 99% homology to the cDNA of a known human gene, vasoactive intestinal peptide receptor 2A. Preliminary data suggests that the *VIPR*2A gene is expressed.
- Identification of numerous expressed sequence tags (ESTs) localized to the 7q region. Since three of the ESTs contain at least two regions with high confidence of homology (~90%), genes in addition to *VIPR*2A may exist in the terminal region of 7q.
- Generation of high-resolution cosmid coverage on human chromosome 5p for the larynx and critical regions identified with Cri-du-Chat syndrome, the most common human terminal-deletion syndrome (in collaboration with Thomas Jefferson University).
- Refinement of the Wolf-Hirschhorn syndrome (WHS) critical region on human chromosome 4p. Using the SCAN program to identify genes likely to contribute to WHS, the project serves as a model for defining the interaction between genomic sequencing and clinical research.
- Collaborative construction of contigs for human chromosome 16, including 1.05 million bases in cosmids through the familial Mediterranean fever (FMF) gene region (with members of the FMF Consortium) and 700,000 bases in P1 clones encompassing the polycystic kidney disease gene (with Integrated Genetics, Inc.).
- Collaborative identification and determination of the complete genomic structure of the Batten's disease gene (with members of the BDG Consortium), the gamma subunit of the human amiloride-sensitive epithelial channel (Liddle's syndrome, with University of Iowa), and the polycystic kidney disease gene (with Integrated Genetics).

- Participation in an international collaborative research consortium that successfully identified the gene responsible for Fanconi anemia type A.

Patents, Licenses, and CRADAs

- Rhett L. Affleck, James N. Demas, Peter M. Goodwin, Jay A. Schecker, Ming Wu, and Richard A. Keller, "Reduction of Diffusional Defocusing in Hydrodynamically Focused Flows by Complexing with a High Molecular Weight Adduct," United States Patent, filed December 1996.
- R.L. Affleck, W.P. Ambrose, J.D. Demas, P.M. Goodwin, M.E. Johnson, R.A. Keller, J.T. Petty, J.A. Schecker, and M. Wu, "Photobleaching to Reduce or Eliminate Luminescent Impurities for Ultrasensitive Luminescence Analysis," United States Patent, S-87, 208, accepted September 1997.
- J.H. Jett, M.L. Hammond, R.A. Keller, B.L. Marrone, and J.C. Martin, "DNA Fragment Sizing and Sorting by Laser-Induced Fluorescence," United States Patent, S.N. 75,001, allowed May 1996.
- James H. Jett, "Method for Rapid Base Sequencing in DNA and RNA with Three Base Labeling," in preparation.
- Development license and exclusive license to LANL's DNA sizing patent obtained by Molecular Technology, Inc., for commercialization of single-molecule detection capability to DNA sizing.

Future Plans

LANL has joined a collaboration with California Institute of Technology and The Institute for Genomic Research to construct a BAC map of the p arm of human chromosome 16 and to complete the sequence of a 20-million–base region of this map.

In its evolving role as part of the new DOE Joint Genome Institute, LANL will continue scaleup activities focused on high-throughput DNA sequencing. Initial targets include genes and DNA regions associated with chromosome structure and function, syntenic break-points, and relevant disease-gene loci.

A joint DNA sequencing center was established recently by LANL at the University of New Mexico. This facility is responsible for determining the DNA sequence of clones constructed at LANL, then returning the data to LANL for analysis and archiving.

Lawrence Berkeley National Laboratory Human Genome Center

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

Michael Palazzolo,* Director, 1996–97

***Now at Amgen, Inc.**

Contact: Mohandas Narla
510/486-7029, Fax: -6746
mohandas_narla@macmail.lbl.gov

Joyce Pfeiffer, Administrative Assistant

http://www-hgc.lbl.gov/GenomeHome.html

Since 1937 the Ernest Orlando Lawrence Berkeley National Laboratory (LBNL) has been a major contributor to knowledge about human health effects resulting from energy production and use. That was the year John Lawrence went to Berkeley to use his brother Ernest's cyclotrons to launch the application of radioactive isotopes in biological and medical research. Fifty years later, Berkeley Lab's Human Genome Center was established.

Now, after another decade, an expansion of biological research relevant to Human Genome Project goals is being carried out within the Life Sciences Division, with support from the Information and Computing Sciences and Engineering divisions. Individuals in these research projects are making important new contributions to the key fields of molecular, cellular, and structural biology; physical chemistry; data management; and scientific instrumentation. Additionally, industry involvement in this growing venture is stimulated by Berkeley Lab's location in the San Francisco Bay area, home to the largest congregation of biotechnology research facilities in the world.

In July 1997 the Berkeley genome center became part of the Joint Genome Institute.

Sequencing

Large-scale genomic sequencing has been a central, ongoing activity at Berkeley Lab since 1991. It has been funded jointly by DOE (for human genome production sequencing and technology development) and the NIH National Human Genome Research Institute [for sequencing the *Drosophila melanogaster* model system, which is carried out in partnership with the University of California, Berkeley (UCB)]. The human genome sequencing area at Berkeley Lab consists of five groups: Bioinstrumentation, Automation, Informatics, Biology, and Development. Complementing these activities is a group in Life Sciences Division devoted to functional genomics, including the transgenics program.

The directed DNA sequencing strategy at Berkeley Lab was designed and implemented to increase the efficiency of genomic sequencing. A key element of the directed approach is maintaining information about the relative positions of potential sequencing templates throughout the entire sequencing process. Thus, intelligent choices can be made about which templates to sequence, and the number of selected templates can be kept to a minimum. More important, knowledge of the interrelationship of sequencing runs guides the assembly process, making it more resistant to difficulties imposed by repeated sequences. As of July 3, 1997, Berkeley Lab had generated 4.4 megabases of human sequence and, in collaboration with UCB, had tallied 7.6 megabases of *Drosophila* sequence.

Instrumentation and Automation

The instrumentation and automation program encompasses the design and fabrication of custom apparatus to facilitate experiments, the programming of laboratory robots to automate repetitive procedures, and the development of (1) improved hardware to extend the applicability range of existing commercial robots and (2) an integrated operating system to control and monitor experiments. Although some discrete instrumentation modules used in the integrated protocols are obtained commercially, LBNL designs its own custom instruments when existing capabilities are inadequate. The instrumentation modules are then integrated into a large system to facilitate large-scale production sequencing. In addition, a significant effort is devoted to improving fluorescence-assay methods, including DNA sequence analysis and mass spectrometry for molecular sizing.

Recent advances in the instrumentation group include DNA Prep machine and Prep Track. These instruments are designed to automate completely the highly repetitive and labor-intensive DNA-preparation procedure to provide higher daily throughput and DNA of consistent quality for sequencing (see Web pages: *http://hgighub.lbl.gov/esd/DNAPrep/TitlePage.html* and *http://hgighub.lbl.gov/esd/repTrackWebpage/preptrack.htm*).

Berkeley Lab's near-term needs are for 960 samples per day of DNA extracted from overnight bacteria growths. The DNA protocol is a modified boil prep prepared in a 96-well

format. Overnight bacteria growths are lysed, and samples are separated from cell debris by centrifugation. The DNA is recovered by ethanol precipitation.

Informatics

The informatics group is focused on hardware and software support and system administration, software development for end sequencing, transposon mapping and sequence template selection, data-flow automation, gene finding, and sequence analysis. Data-flow automation is the main emphasis. Six key steps have been identified in this process, and software is being written and tested to automate all six. The first step involves controlling gel quality, trimming vector sequence, and storing the sequences in a database. A program module called Move-Track-Trim, which is now used in production, was written to handle these steps. The second through fourth steps in this process involve assembling, editing, and reconstructing P1 clones of 80,000 base pairs from 400-base traces. The fifth step is sequence annotation, and the sixth is data submission.

Annotation can greatly enhance the biological value of these sequences. Useful annotations include homologies to known genes, possible gene locations, and gene signals such as promoters. LBNL is developing a workbench for automatic sequence annotation and annotation viewing and editing. The goal is to run a series of sequence-analysis tools and display the results to compare the various predictions. Researchers then will be able to examine all the annotations (for example, genes predicted by various gene-finding methods) and select the ones that look best.

Nomi Harris developed Genotator, an annotation workbench consisting of a stand-alone annotation browser and several sequence-analysis functions. The back end runs several gene finders, homology searches (using BLAST), and signal searches and saves the results in ".ace" format. Genotator thus automates the tedious process of operating a dozen different sequence-analysis programs with many different input and output formats. Genotator can function via command-line arguments or with the graphical user interface (*http://www-hgc.lbl.gov/inf/annotation.html*).

Progress to Date

Chromosome 5

Over the last year, the center has focused its production genomic sequencing on the distal 40 megabases of the human chromosome 5 long arm. 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. Results to date include:

- 40-megabase nonchimeric map containing 82 yeast artificial chromosomes (YACs) in the chromosome 5 distal long arm.
- 20-megabase contig map in the region of 5q23-q33 that contains 198 P1s, 60 P1 artificial chromosomes, and 495 bacterial artificial chromosomes (BACs) linked by 563 sequenced tagged sites (STSs) to form contigs.
- 20-megabase bins containing 370 BACs in 74 bins in the region of 5q33-q35.

Chromosome 21

An early project in the study of Down syndrome (DS), which is characterized by chromosome 21 trisomy, constructed a high-resolution clone map in the chromosome 21 DS region to be used as a pilot study in generating a contiguous gene map for all of chromosome 21. This project has integrated P1 mapping efforts with transgenic studies in the Life Sciences Division. P1 maps provide a suitable form of genomic DNA for isolating and mapping cDNA.

- 186 clones isolated in the major DS region of chromosome 21 comprising about 3 megabases of genomic DNA extending from D21S17 to ETS2. Through cross-hybridization, overlapping P1s were identified, as well as gaps between two P1 contigs, and transgenic mice were created from P1 clones in the DS region for use in phenotypic studies.

Transgenic Mice

One of the approaches for determining the biological function of newly identified genes uses YAC transgenic mice. Human sequence harbored by YACs in transgenic mice has been shown to be correctly regulated both temporally and spatially. A set of nonchimeric overlapping YACs identified from the 5q31 region has been used to create transgenic mice. This set of transgenic mice, which together harbor 1.5 megabases of human sequence, will be used to assess the expression pattern and potential function of putative genes discovered in the 5q31 region. Additional mapping and sequencing are under way in a region of human chromosome 20 amplified in certain breast tumor cell lines.

Resource for Molecular Cytogenetics

Divining landmarks for human disease amid the enormous plain of the human genetic map is the mission of an ambitious partnership among the Berkeley Lab; University of California, San Francisco; and a diagnostics company. The collaborative Resource for Molecular Cytogenetics is charting a course toward important sites of biological interest on the 23 pairs of human chromosomes (*http://rmc-www.lbl.gov*).

The Resource employs the many tools of molecular cytogenetics. The most basic of these tools, and the cornerstone of the Resource's portfolio of proprietary technology, is a method generally known as "chromosome painting," which uses a technique referred to as fluorescence in situ hybridization or FISH. This technology was invented by LBNL Resource leaders Joe Gray and Dan Pinkel.

A technology to emerge recently from the Resource is known as "Quantitative DNA Fiber Mapping (QDFM)." High-resolution human genome maps in a form suitable for DNA sequencing traditionally have been constructed by various methods of fingerprinting, hybridization, and identification of overlapping STSs. However, these techniques do not readily yield information about sequence orientation, the extent of overlap of these elements, or the size of gaps in the map. Ulli Weier of the Resource developed the QDFM method of physical map assembly that enables the mapping of cloned DNA directly onto linear, fully extended DNA molecules. QDFM allows unambiguous assembly of critical elements leading to high-resolution physical maps. This task now can be accomplished in less than 2 days, as compared with weeks by conventional methods. QDFM also enables detection and characterization of gaps in existing physical maps—a crucial step toward completing a definitive human genome map.

University of Washington Genome Center
Department of Medicine
Box 352145
Seattle, WA 98195

Maynard Olson, Director
206/685-7366, Fax: -7344
mvo@u.washington.edu

http://www.genome.washington.edu

The Human Genome Project soon will need to increase rapidly the scale at which human DNA is analyzed. The ultimate goal is to determine the order of the 3 billion bases that encode all heritable information. During the 20 years since effective methods were introduced to carry out DNA sequencing by biochemical analysis of recombinant-DNA molecules, these techniques have improved dramatically. In the late 1970s, segments of DNA spanning a few thousand bases challenged the capacity of world-class sequencing laboratories. Now, a few million base pairs per year represent state-of-the-art output for a single sequencing center.

However, the Human Genome Project is directed toward completing the human sequence in 5 to 10 years, so the data must be acquired with technology available now. This goal, while clearly feasible, poses substantial organizational and technical challenges. Organizationally, genome centers must begin building data-production units capable of sustained, cost-effective operation. Technically, many incremental refinements of current technology must be introduced, particularly those that remove impediments to increasing the scale of DNA sequencing. The University of Washington (UW) Genome Center is active in both areas.

Production Sequencing

Both to gain experience in the production of high-quality, low-cost DNA sequence and to generate data of immediate biological interest, the center is sequencing several regions of human and mouse DNA at a current throughput of 2 million bases per year. This "production sequencing" has three major targets: the human leukocyte antigen (HLA) locus on human chromosome 6, the mouse locus encoding the alpha subunit of T-cell receptors, and an "anonymous" region of human chromosome 7.

The HLA locus encodes genes that must be closely matched between organ donors and organ recipients. This sequence data is expected to lead to long-term improvements in the ability to achieve good matches between unrelated organ donors and recipients.

The mouse locus that encodes components of the T-cell–receptor family is of interest for several reasons. The locus specifies a set of proteins that play a critical role in cell-mediated immune responses. It provides sequence data that will help in the design of new experimental approaches to the study of immunity in mice—one of the most important experimental animals for immunological research. In addition, the locus will provide one of the first large blocks of DNA sequence for which both human and mouse versions are known.

Human-mouse sequence comparisons provide a powerful means of identifying the most important biological features of DNA sequence because these features are often highly conserved, even between such biologically different organisms as human and mouse. Finally, sequencing an "anonymous" region of human chromosome 7, a region about which little was known previously, provides experience in carrying out large-scale sequencing under the conditions that will prevail throughout most of the Human Genome Project.

Technology for Large-Scale Sequencing

In addition to these pilot projects, the UW Genome Center is developing incremental improvements in current sequencing technology. A particular focus is on enhanced computer software to process raw data acquired with automated laboratory instruments that are used in DNA mapping and sequencing. Advanced instrumentation is commercially available for determining DNA sequence via the "four-color–fluorescence method," and this instrumentation is expected to carry the main experimental load of the Human Genome Project. Raw data produced by these instruments, however, require extensive processing before they are ready for biological analysis.

Large-scale sequencing involves a "divide-and-conquer" strategy in which the huge DNA molecules present in human cells are broken into smaller pieces that can be propagated by recombinant-DNA methods. Individual analyses ultimately are carried out on segments of less than 1000 bases. Many such analyses, each of which still contains numerous errors, must be melded together to obtain finished sequence. During the melding, errors in individual analyses must be recognized and corrected. In typical large-scale sequencing projects, the results of thousands of analyses are melded to produce highly accurate sequence (less than one error in 10,000 bases) that is continuous in blocks of 100,000 or more bases. The UW Genome Center is playing a major role in developing software that allows this process to be carried out automatically with little need for expert intervention. Software developed in the UW center is used in more than 50 sequencing laboratories around the world, including most of the large-scale sequencing centers producing data for the Human Genome Project.

High-Resolution Physical Mapping

The UW Genome Center also is developing improved software that addresses a higher-level problem in large-scale sequencing. The starting point for large-scale sequencing typically is a recombinant-DNA molecule that allows propagation of a particular human genomic segment spanning 50,000 to 200,000 bases. Much effort during the last decade has gone into the physical mapping of such molecules, a process that allows huge regions of chromosomes to be defined in terms of sets of overlapping recombinant-DNA molecules whose precise positions along the chromosome are known. However, the precision required for knowing relationships of recombinant-DNA molecules derived from neighboring chromosomal portions increases as the Human Genome Project shifts its emphasis from mapping to sequencing.

High-resolution maps both guide the orderly sequencing of chromosomes and play a critical role in quality control. Only by mapping recombinant-DNA molecules at high resolution can subtle defects in particular molecules be recognized. Such defective human DNA sources, which are not faithful replicas of the human genome, must be weeded out before sequencing can begin. The UW Genome Center has a major program in high-resolution physical mapping which, like the work on sequencing itself, uses advanced computing tools. The center is producing maps of regions targeted for sequencing on a just-in-time basis. These highly detailed maps are proving extremely valuable in facilitating the production of high-quality sequence.

Ultimate Goal

Although many challenges currently posed by the Human Genome Project are highly technical, the ultimate goal is biological. The project will deliver immense amounts of high-quality, continuous DNA sequence into publicly accessible databases. These data will be annotated so that biologists who use them will know the most likely positions of genes and have convenient access to the best available clues about the probable function of these genes. The better the technical solutions to current challenges, the better the center will be able to serve future users of the human genome sequence.

Genome Database
Johns Hopkins University
2024 E. Monument Street
Baltimore, MD 21205-2236

David Kingsbury, Director, 1993–97*

*Now at Chiron Pharmaceuticals, Emeryville, California

Stanley Letovsky, Informatics Director
letovsky@gdb.org

Robert Cottingham, Operations Director
bc@gdb.org

Telephone for both: 410/955-9705
Fax for both: 410/614-0434

http://www.gdb.org

The release of Version 6 of the Genome Database (GDB) in January 1996 signaled a major change for both the scientific community and GDB staff. GDB 6.0 introduced a number of significant improvements over previous versions of GDB, most notably a revised data representation for genes and genomic maps and a new curatorial model for the database. These new features, along with a remodeled database structure and new schema and user interface, provide a resource with the potential to integrate all scientific information currently available on human genomics. GDB rapidly is becoming the international biomedical research community's central source for information about genomic structure, content, diversity, and evolution.

A New Data Model

Inherent in the underlying organization of information in GDB is an improved model for genes, maps, and other classes of data. In particular, genomic segments (any named region of the genome) and maps are being expanded regularly. New segment types have been added to support the integration of mapping and sequencing data (for example, gene elements and repeats) and the construction of comparative maps (syntenic regions). New map types include comparative maps for representing conserved syntenies between species and comprehensive maps that combine data from all the various submitted maps within GDB to provide a single integrated view of the genome. Experimental observations such as order, size, distance, and chimerism are also available.

Through the World Wide Web, GDB links its stored data with many other biological resources on the Internet. GDB's External Link category is a growing collection of cross-references established between GDB entities and related information in other databases. By providing a place for these cross-references, GDB can serve as a central point of inquiry into technical data regarding human genomics.

Direct Community Data Submission and Curation

Two methods for data submission are in use. For individuals submitting small amounts of data, interactive editing of the database through the Web became available in April 1996, and the process has undergone several simplifications since that time. This continues to be an area of development for GDB because all editing must take place at the Baltimore site, and Internet connections from outside North America may be too slow for interactive editing to be practical. Until these difficulties are resolved, GDB encourages scientists with limited connectivity to Baltimore to submit their data via more traditional means (e-mail, fax, mail, phone) or to prepare electronic submissions for entry by the data group on site.

For centers submitting large quantities of data, GDB developed an electronic data submission (EDS) tool, which provides the means to specify login password validation and commands for inserting and updating data in GDB. The EDS syntax includes a mechanism for relating a center's local naming conventions to GDB objects. Data submitted to GDB may be stored privately for up to 6 months before it automatically becomes public. The database is programmed to enforce this Human Genome Project policy. Detailed specifications of GDB's EDS syntax and other submission instructions are available (EDS prototype, *http://www.gdb.org/eds*).

Since the EDS system was implemented, GDB has put forth an aggressive effort to increase the amount of data stored in the database. Consequently, the database has grown tremendously. During 1996 it grew from 1.8 to 6.7 gigabytes.

To provide accountability regarding data quality, the shift to community curation introduced the idea that individuals and laboratories own the data they submit to GDB and that other researchers cannot modify it. However, others should be able to add information and comments, so an additional feature is the commu-nity's ability to conduct electronic online public discussions by annotating the

database submissions of fellow researchers. GDB is the first database of its kind to offer this feature, and the number of third-party annotations is increasing in the form of editorial commentary, links to literature citations, and links to other databases external to GDB. These links are an important part of the curatorial process because they make other data collections available to GDB users in an appropriate context.

Improved Map Representation and Querying

Accompanying the release of GDB 6.0, the program Mapview creates graphical displays of maps. Mapview was developed at GDB to display a number of map types (cytogenetic, radiation hybrid, contig, and linkage) using common graphical conventions found in the literature. Mapview is designed to stand alone or to be used in conjunction with a Web browser such as Netscape, thereby creating an interactive graphical display system. When used with Netscape, Mapview allows the user to retrieve details about any displayed map object.

Maps are accessed through the query form for genomic segment and its subclasses via a special program that allows the user to select whole maps or slices of maps from specific regions of interest and to query by map type. The ability to browse maps stored in GDB or download them in the background was also incorporated into GDB 6.0.

GDB stores many maps of each chromosome, generated by a variety of mapping methods. Users who are interested in a region, such as the neighborhood of a gene or marker, will be able to see all maps that have data in that region, whether or not they contain the desired marker. To support database querying by region of interest, integrated maps have been developed that combine data from all the maps for each chromosome. These are called Comprehensive Maps.

Queries for all loci in a region of interest are processed against the comprehensive maps, thereby searching all relevant maps. Comprehensive maps are also useful for display purposes because they organize the content of a region by class of locus (e.g., gene, marker, clone) rather than by data source. This approach yields a much less complex presentation than an alignment of numerous primary maps. Because such information as detailed orders, order discrepancies between maps, and nonlinear metric relations between maps is not always captured in the comprehensive maps, GDB continues to provide access to aligned displays of primary maps.

A Variety of Searching Strategies

Recognizing the eclectic user commu-nity's need to search data and formulate queries, GDB offers a spectrum of simple to complex search strategies. In addition, direct programming access is available using either GDB's object query language to the Object Broker software layer or standard query language to the underlying Sybase relational database.

Querying by Object Directly from GDB's Home Page

The simplest methods search for objects according to known GDB accession numbers; sequence database–accession numbers; specified names, including wildcard symbols that will automatically match synonyms and primary names; and keywords contained anywhere in the text.

Querying by Region of Interest

A region of interest can be specified using a pair of flanking markers, which can be cytogenetic bands, genes, amplimers (sequence tagged sites), or any other mapped objects. Given a region of interest, the comprehensive maps are searched to find all loci that fall within them. These loci can be displayed in a table, graphically as a slice through a comprehensive map, or as slices through a chosen set of primary maps. A comprehensive map slice shows all loci in the region, including genes, expressed sequence tags (ESTs), amplimers, and clones. A region also can be specified as a neighborhood around a single marker of interest.

Results of queries for genes, amplimers, ESTs, or clones can be displayed on a GDB comprehensive map. Results are spread across several chromosomes displayed in Mapview. A query for all the PAX genes (specified as symbol = PAX* on the gene query form) retrieves genes on multiple chromosomes. Double-clicking on one of these genes brings up detailed gene information via the Web browser.

Querying by Polymorphism

GDB contains a large number of polymorphisms associated with genes and other markers. Queries can be constructed for a particular type of marker (e.g., gene, amplimer, clone), polymorphism (i.e., dinucleotide repeat), or level of heterozygosity. These queries can be combined with positional queries to find, for example, polymorphic amplimers in a region bounded by flanking markers or in a particular chromosomal band. If desired, the retrieved markers can be viewed on a comprehensive map.

Work in Progress

Mapview 2.3

Mapview 2.1, the next generation of the GDB map viewer, was released in March 1997. The latest version, Mapview 2.3, is available in all common computing environments because it is written in the Java programming language. Most important, the new viewer can display multiple aligned maps side by side in the window, with alignment lines indicating common markers in neighboring maps. As before, users can select individual markers to retrieve more information about them from the database.

GDB developers have entered into a collaborative relationship with other members of the bioWidget Consortium so the Java-based alignment viewer will become part of a collection of freely available software tools for displaying biological data (*http://goodman.jax.org/projects/biowidgets/consortium*).

Future plans for Mapview include providing or enhancing the ability to generate manuscript-ready Postscript map images, highlight or modify the display of particular classes of map objects based on attribute values, and requery for additional information.

Variation

Since its inception, GDB has been a repository for polymorphism data, with more than 18,000 polymorphisms now in GDB. A collaboration has been initiated with the Human Gene Mutation Database (HGMD) based in Cardiff, Wales, and headed by David Cooper and Michael Krawczak. HGMD's extensive collection of human mutation data, covering many disease-causing loci, includes sequence-level mutation characterizations. This data set will be included in GDB and updated from HGMD on an ongoing basis. The HGMD team also will provide advice on GDB's representation of genetic variation, which is being enhanced to model mutations and polymorphisms at the sequence level. These modifications will allow GDB to act as a repository for single-nucleotide polymorphisms, which are expected to be a major source of information on human genetic variation in the near future.

Mouse Synteny

Genomic relationships between mouse and man provide important clues regarding gene location, phenotype, and function. One of GDB's goals is to enable direct comparisons between these two organisms, in collaboration with the Mouse Genome Database at Jackson Laboratory. GDB is making additions to its schema to represent this information so that it can be displayed graphically with Mapview. In addition, algorithmic work is under way to use mapping data to automatically identify regions of conserved synteny between mouse and man. These algorithms will allow the synteny maps to be updated regularly. An important application of comparative mapping is the ability to predict the existence and location of unknown human homologs of known, mapped mouse genes. A set of such predictions is available in a report at the GDB Web site, and similar data will be available in the database itself in the spring of 1998.

Collaborations

GDB is a participant in the Genome Annotation Consortium (GAC) project, whose goal is to produce high-quality, automatic annotation of genomic sequences (*http://compbio.ornl.gov/CoLab*). Currently, GDB is developing a prototype mechanism to transition from GDB's Mapview display to the GAC sequence-level browser over common genome regions. GAC also will establish a human genome reference sequence that will be the base against which GDB will refer all polymorphisms and mutations. Ultimately, every genomic object in GDB should be related to an appropriate region of the reference sequence.

Sequencing Progress

The sequencing status of genomic regions now can be recorded in GDB. Based on submissions to sequence databases, GAC will determine genomic regions that have been completed. GDB also will be collaborating with the European Bioinformatics Institute, in conjunction with the international Human Genome Organisation (HUGO), to maintain a single shared Human Sequence Index that will record commitments and status for sequencing clones or regions. As a result, the sequencing status of any region can be displayed alongside other GDB mapping data.

Outreach

The Genome Database continues to seek direct community feedback and interact with the broader science community via various sources:

- International Scientific Advisory Committee meets annually to offer input and advice.
- Quarterly Review Committee confers frequently with the staff to track GDB progress and suggest change.
- HUGO nomenclature, chromosome, and other editorial committees have specialized functions within GDB, providing official names and consensus maps and ensuring the high quality of GDB's content.

Copies of GDB are available worldwide from ten mirror sites (nodes), and GDB staff members meet annually with node managers.

Genome Sequence DataBase
1800 Old Pecos Trail, Suite A
Santa Fe, NM 87505

Peter Schad, Vice-President
Bioinformatics and Biotechnology
505/995-4447, Fax: -4432
cnc@ncgr.org

Carol Harger
GSDB Manager
505/982-7840, Fax: -7690
cah@ncgr.org

http://www.ncgr.org

The National Center for Genome Resources (NCGR) is a not-for-profit organization created to design, develop, support, and deliver resources in support of public and private genome and genetic research. To accomplish these goals, NCGR is developing and publishing the Genome Sequence DataBase (GSDB) and the Genetics and Public Issues (GPI) program.

NCGR is a center to facilitate the flow of information and resources from genome projects into both public and private sectors. A broadly based board of governors provides direction and strategy for the center's development.

NCGR opened in Santa Fe in July 1994, with its initial bioinformatics work being developed through a cooperative 5-year agreement with the Department of Energy funded in July 1995. Committed to serving as a resource for all genomic research, the center works collaboratively with researchers and seeks input from users to ensure that tools and projects under development meet their needs.

Genome Sequence DataBase

GSDB is a relational database that contains nucleotide sequence data and its associated annotation from all known organisms (*http://www.ncgr.org/gsdb*). All data are freely available to the public. The major goals of GSDB are to provide the support structure for storing sequence data and to furnish useful data-retrieval services.

GSDB adheres to the philosophy that the database is a "community-owned" resource that should be simple to update to reflect new discoveries about sequences. A corollary to this is GSDB's conviction that researchers know their areas of expertise much better than a database curator and, therefore, they should be given ownership and control over the data they submit to the database. The true role of the GSDB staff is to help researchers submit data to and retrieve data from the database.

GSDB Enhancements

During 1996, GSDB underwent a major renovation to support new data types and concepts that are important to genomic research. Tables within the database were restructured, and new tables and data fields were added. Some key additions to GSDB include the support of data ownership, sequence alignments, and discontiguous sequences.

The concept of data ownership is a cornerstone to the functioning of the new GSDB. Every piece of data (e.g., sequence or feature) within the database is owned by the submitting researcher, and changes can be made only by the data owner or GSDB staff. This implementation of data ownership provides GSDB with the ability to support community (third-party) annotation—the addition of annotation to a sequence by other community researchers.

A second enhancement of GSDB is the ability to store and represent sequence alignments. GSDB staff has been constructing alignments to several key sequences including the env and pol (reverse transcriptase) genes of the HIV genome, the complete chromosome VIII of Saccharomyces cerevisiae, and the complete genome of Haemophilus influenzae. These alignments are useful as possible sites of biological interest and for rapidly identifying differences between sequences.

A third key GSDB enhancement is the ability to represent known relationships of order and distance between separate individual pieces of sequence. These sets of sequences and their relative positions are grouped together as a single discontiguous sequence. Such a sequence may be as simple as two primers that define the ends of a sequence tagged site (STS), it may comprise all exons that are part of a single gene, or it may be as complex as the STS map for an entire chromosome.

GSDB staff has constructed discontigu-ous sequences for human chromosomes 1 through 22 and X that include markers from Massachusetts Institute of Technology–Whitehead Institute STS maps and from the Stanford Human Genome Center. The set of 2000 STS markers for chromosome X, which were mapped recently by Washington University at St. Louis, also have been added to chromosome X. About 50 genomic sequences have been added to the chromosome 22 map by determining their overlap with STS markers. Genomic sequences are being added to all the chromosomes as their overlap with the STS markers is determined. These discontigu-ous sequences can be retrieved easily and viewed via their sequence names using

the GSDB Annotator. Sequence names follow the format of HUMCHR#MP, where # equals 1 through 22 or X.

GSDB staff also has utilized discontigu-ous sequences to construct maps for maize and rice. The maize discontiguous sequences were constructed using markers from the University of Missouri, Columbia. Markers for the rice discontiguous sequence were obtained from the Rice Genome Database at Cornell University and the Rice Genome Research Project in Japan.

New Tools

As a result of the major GSDB renovation, new tools were needed for submitting and accessing database data. Annotator was developed as a graphical interface that can be used to view, update, and submit sequence data (*http://www.ncgr.org/gsdb/beta.html*). Maestro, a Web-based interface, was developed to assist researchers in data retrieval (*http://www.ncgr.org/gsdb/maestrobeta.html*). Although both these tools currently are available to researchers, GSDB is continuing development to add increased capabilities.

Annotator displays a sequence and its associated biological information as an image, with the scale of the image adjustable by the user. Additional information about the sequence or an associate biological feature can be obtained in a pop-up window. Annotator also allows a user to retrieve a sequence for review, edit existing data, or add annotation to the record. Sequences can be created using Annotator, and any sequences created or edited can be saved either to a local file for later review and further editing or saved directly to the database.

Correct database structures are important for storing data and providing the research community with tools for searching and retrieving data. GSDB is making a concerted effort to expand and improve these services. The first generation of the Maestro query tool is available from the GSDB Web pages. Maestro allows researchers to perform queries on 18 different fields, some of which are queryable only through GSDB, for example, D segment numbers from the Genome Database at Johns Hopkins University in Baltimore.

Additionally, Maestro allows queries with mixed Boolean operators for a more refined search. For example, a user may wish to compare relatively long mouse and human sequences that do not contain identified coding regions. To obtain all sequences meeting these criteria, the scientific name field would be searched first for “Mus musculus” and then for “Homo sapiens” using the Boolean term “OR.” Then the sequence-length filter could be used to refine the search to sequences longer than 10,000 base pairs. To exclude sequences containing identified coding-region features, the “BUT NOT” term can be used with the Feature query field set equal to “coding region.”

With Maestro, users can view the list of search matches a few at a time and retrieve more of the list as needed. From the list, users can select one or several sequences according to their short descriptions and review or download the sequence information in GIO, FASTA, or GSDB flatfile format.

Future Plans

Although most pieces necessary for operation are now in place, GSDB is still improving functionality and adding enhancements. During the next year GSDB, in collaboration with other researchers, anticipates creating more discontiguous sequence maps for several model organisms, adding more functionality to and providing a Web-based submission tool and tool kit for creating GIO files.

Microbial Genome Web Page

NCGR also maintains informational Web pages on microbial genomes. These pages, created as a community reference, contain a list of current or completed eubacterial, Archaeal, and eukaryotic genome sequencing projects. Each main page includes the name of the organism being sequenced, sequencing groups involved, background information on the organism, and its current location on the Carl Woese Tree of Life. As the Microbial Genome Project progresses, the pages will be updated as appropriate.

Genetics and Public Issues Program

GPI serves as a crucial resource for people seeking information and making decisions about genetics or genomics (*http://www.ncgr.org/gpi*). GPI develops and provides information that explains the ethical, legal, policy, and social relevance of genetic discoveries and applications.

To achieve its mission, GPI has set forth three goals: (1) preparation and development of resources, including careful delineation of ethical, legal, policy, and social issues in genetics and genomics; (2) dissemination of genetic information targeted to the public, legal and health professionals, policymakers, and decision makers; and (3) creation of an information network to facilitate interaction among groups.

GPI delivers information through four primary vehicles: online resources, conferences, publications, and educational programs. The GPI program maintains a continually evolving World Wide Web site containing a range of material freely accessible over the Internet.

Index to Principal and Coinvestigators Listed in Abstracts

D

E

F

G

H

N

O

P

R

S

T

U

V

W