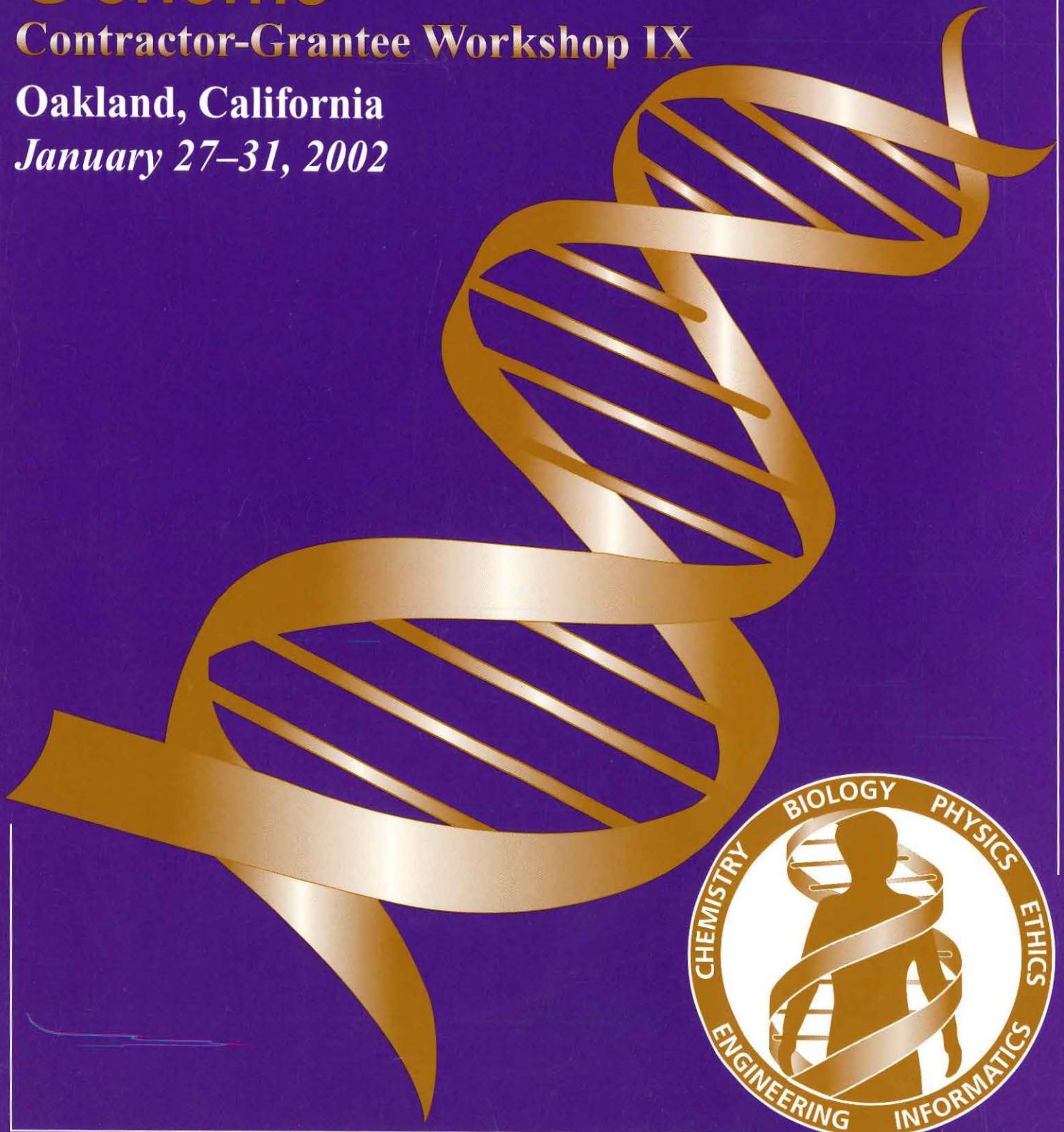
DOE/SC-0044

DOE Genome

Contractor-Grantee Workshop IX

Oakland, California

January 27–31, 2002



Contents

Poster Number | Page

Poster Number | Page

Poster Number | Page

Poster Number | Page

Poster Number

Page

Poster Number | Page

Poster Number | Page

Introduction to Contractor-Grantee Workshop IX

Welcome to the Ninth Contractor-Grantee Workshop sponsored by the Department of Energy (DOE) genomics programs. This workshop provides a unique opportunity for DOE genome investigators to discuss and share the successes, problems, and challenges of their research as well as new resources and software capabilities. The meeting also provides scientists and administrative staff with an overview of the program's progress and content, a chance to assess the impact of new technologies, and, perhaps most important, a forum for initiating new collaborations. We hope you will take full advantage of the opportunities offered by this meeting and include a visit to the DOE Joint Genome Institute's Production Genomics Facility in Walnut Creek.

The 201 abstracts in this booklet describe the most recent activities and accomplishments of grantees and contractors funded by DOE's human genome and microbial genome programs, as well as the more recent Genomes to Life initiative. We also have included talks from invited guests who will discuss related efforts and opportunities for the biology enabled by genome research. All genome projects funded by the Biological and Environmental Research program will be represented at poster sessions. Plan to meet with the researchers who make this program a success and take advantage of all the formal and informal opportunities for discussion and exchange of information available at this workshop.

With the draft human sequence gradually maturing into high-quality, "Bermuda-Standard" finished sequence, the full complement of genes eventually will be identified. Completion of gene inventories for the human, several model organisms, and an increasing number of microbes has forever changed the practice and power of biology. Simple descriptions of genes and their most evident actions are no longer adequate. The overarching goal now is to achieve effective quantitative, and hence testable, predictive models of cells and their many constituent processes. A major challenge lies in the large number of complex macromolecular machines that manage and transact the many processes and adaptations of a cell.

With respect to the broad range of interesting potential target organisms, the DOE role can be comprehensive for some and enabling for many others. Many enabling resources and technologies initially sponsored by DOE in the Human Genome Project have since enjoyed broad usage. These include BAC resources for mapping and sequencing, thermosequenases, improved fluorescent dye labels, and capillary-based DNA sequencers. The DOE Production Genomics Facility represents another enabling resource, advancing the understanding of genomes of many species by generating annotated draft genome sequence in a cost-effective process. Many other emerging capabilities evident in the newer instrumentation and informatics technologies supported by DOE promise to bring high-throughput efficiencies to many aspects of functional analysis.

The Genomes to Life initiative, begun in FY 2002, has a comprehensive agenda developed by experts in diverse disciplines. Its initial targets are microbes having beneficial roles in generating clean energy, mitigating global climate change, restoring the environment, and neutralizing bioterrorism threats. With only thousands of genes directing microbial cell actions, the effective modeling of multicomponent

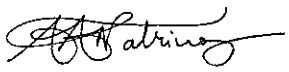
protein complexes and their interactions with the environment is more tractable than in higher organisms.

Within the Human Genome Project, DOE is finishing chromosomes 5, 16, and 19, representing around 10% (or 250 million base pairs) of the euchromatin in the human genome. The mechanisms of gene regulation in these three chromosomes will be of continuing DOE interest. Elucidation of these mechanisms benefits from the sequencing of several model organisms or their genomic regions syntenic with human chromosomes 5, 16, and 19. Particular genes of interest are those mediating individual susceptibilities to environmental toxins and ionizing radiation.

Although many challenges lie ahead, particularly in anticipating and preparing for the post Human Genome Project era, we are more optimistic than ever about the success of this grand project and its many contributions to science and society. Yet we cannot afford to be complacent, and workshop presenters on ethical, legal, and social implications (ELSI) will remind and challenge all of us that science has societal impacts that we must confront. ELSI implications are not just research topics for ELSI investigators but real-life issues that need to be considered in the context of all genome research and with the active participation of all involved scientists.

We look forward to a very interesting and productive meeting and offer our sincere thanks to all the organizers and to you, the scientists, whose vision and efforts continue to realize the promise of genome research.

Sincerely,



Ari Patrinos
Associate Director of Science for Biological and Environmental Research
Office of Biological and Environmental Research
U.S. Department of Energy
genome@science.doe.gov

GENOMES to LIFE
BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES
U.S. DEPARTMENT OF ENERGY
INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS

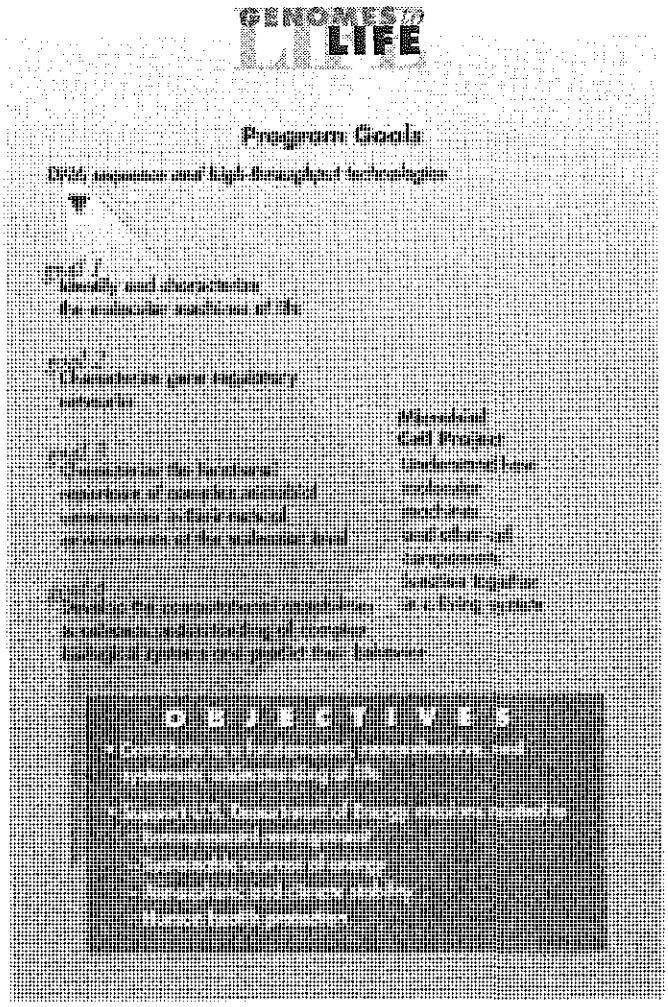
Program Overview

Office of Science

Built on the continuing successes of international genome-sequencing projects, the Genomes to Life program will take the logical next step: a quest to understand the composition and function of the biochemical networks and pathways that carry out the essential processes of living organisms. The roadmap published in April 2001 sets forth an aggressive 10-year plan designed to exploit high-throughput genomic strategies and centered around the four major goals outlined in the chart at right.

The Genomes to Life program reflects the fundamental change now occurring in the way biologists think about biology, a perspective that is a logical and compelling product of the Human Genome Project (HGP). The new program will build on HGP achievements, both by exploiting its data and by extending its paradigm of comprehensive, whole-genome biology to the next level. This approach ultimately will enable an integrated and predictive understanding of biological systems—an understanding that will offer insights into how both microbial and human cells respond to environmental changes. The applications of this next level of understanding will be revolutionary.

The current state-of-the-art instrumentation and computation enable and encourage the immediate establishment of this ambitious and far-reaching program. The strategic alliance created between DOE's offices of Advanced Scientific Computing Research (ASCR) and Biological and Environmental Research (BER) will develop the infrastructure to meet these challenges. Concurrent technology development also will be needed to reach all goals within the next decade. Substantial efforts will be devoted, for example, to improving technologies for characterizing proteins and protein complexes, localizing them in cells and tissues, carrying out high-throughput functional assays of complete cellular protein inventories,



and sequencing and analyzing microbial DNA taken from natural environments.

The Genomes to Life program complements and augments the Department of Energy's (DOE) Microbial Cell Project, launched in FY 2001. The goal of this established project is to collect, analyze, and integrate data on individual microbes in an effort to understand how cellular components function together to create living systems, particularly those with capabilities of interest to DOE.

DOE is strongly positioned to make major contributions to the scientific advances promised by the biology of the 21st century. Strengths of DOE's national laboratories include major facilities for DNA sequencing and molecular structure characterization, high-performance computing resources, the expertise and infrastructure for technology development, and a legacy of productive multidisciplinary research essential for such an ambitious and complex program. In the effort to understand biological systems, these assets and the Genomes to Life program will complement and fundamentally enable the capabilities and efforts of the National Institutes of Health, the National Science Foundation, and other agencies and institutions around the world.

Genomes to Life Program

GTL was developed in response to a 1999 charge by the DOE Office of Science to the Biological and Environmental Research Advisory Comittee to define DOE's potential roles in post-HGP science. The resulting August 2000 report, *Bringing the Genome to Life*, set forth recommendations that led to the roadmap published in April 2001. The FY 2002 budget for GTL is $19.5 million.

DOE Contacts

BER, Marvin Frazier
301/903-5468, Fax: /903-8521
marvin.frazier@science.doe.gov

ASCR, Gary Johnson
970/225-3794, Fax: /223-1415
garyj@er.doe.gov

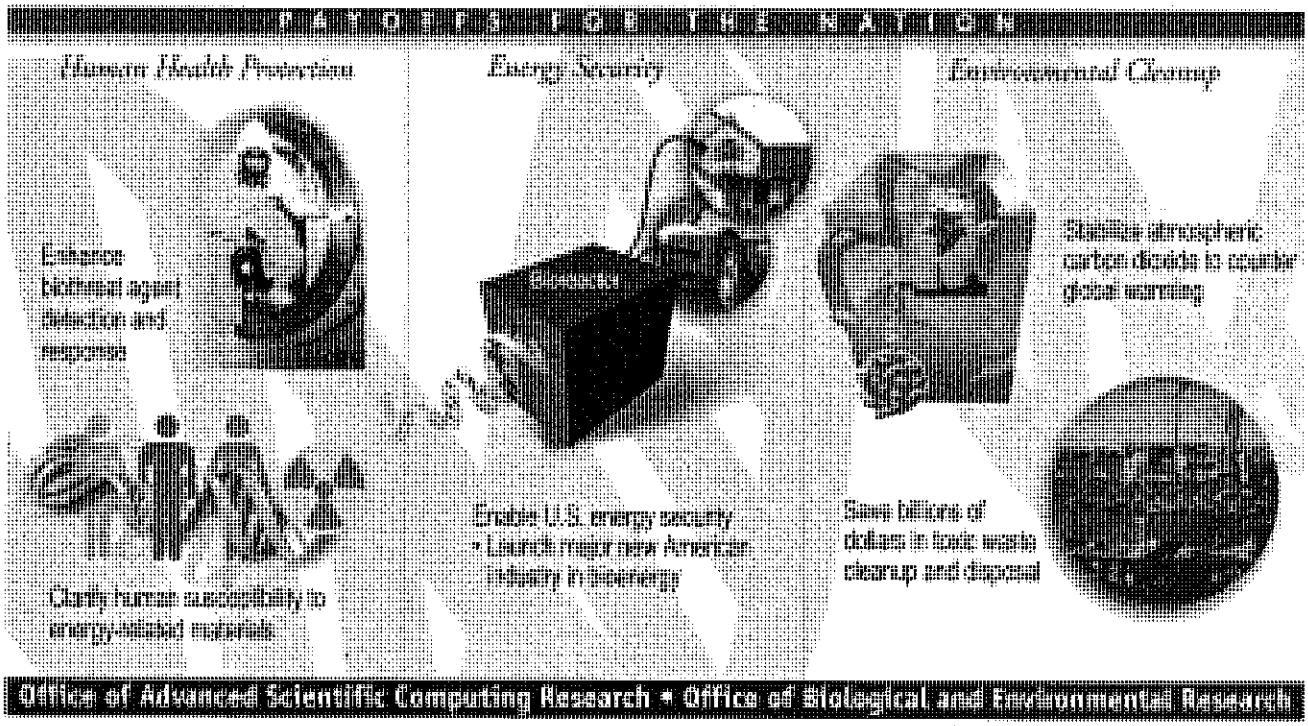
GTL Publications

Documents, meeting reports, and image gallery are downloadable via the Web:

- DOEGenomesToLife.org

Requests for future publications:

- Human Genome Management Information System
  865/576-6669, Fax: /574-9888
  mansfieldbk@ornl.gov



Sequencing

1. The US DOE Joint Genome Institute's High Throughput Production Sequencing Program

Susan Lucas, Tijana Glavina, Jamie Jett, Lyle Probst, Andrea Aerts, Nathan Bunker, Sanjay Israni, Astrid Terry, John C. Detter, Sam Pitluck, Heather Kimball, Yunian Lou, Martin Pollard, Anne Olsen, Chris Elkin, Paul Richardson, Dan Rokhsar, Paul Predki, Elbert Branscomb, Trevor Hawkins, and the JGI Sequencing Team
U.S. DOE Joint Genome Institute, Walnut Creek, CA 94598
lucas11@llnl.gov

In May 2001, the Department of Energy's Joint Genome Institute (JGI) Production Genomics Facility (PGF) automated the use of rolling circle amplification (RCA) as a way to amplify plasmids for high throughput sequencing. With this new approach we are able to produce uniform amounts of template DNA resulting in high quality sequencing results. In addition, this new process has reduced the number of steps for template production, as compared to our previous magnetic bead plasmid preparation, (SPRI). These processes were automated using various liquid transfer robots and a series of quality controls were put in place with each process to track the quality at various stages. The changes have resulted in a simple process that allows for careful monitoring of quality and a significant cost savings in terms of number of steps, time, and people used to produce high quality results. Since May, the US DOE JGI has been concentrating on using this new production line to complete the sequencing of the human chromosomes 5, 16 and 19 as well as several other large genomes such as *Fugu rupies* and *Ciona intestinalis*. The PGF is currently building a microbial sequencing program and that will sequence several microbe genomes throughout FY02. In December, the PGF will install new technology bringing 21 Molecular Dynamic MegaBACE 4000 instruments. This 384 well capillary electrophoresis sequencer will increase sequencing throughput by 40% and enable the JGI institute to ramp from 150 to 250 384 well plates.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, the Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098, and the Los Alamos National Laboratory under contract No. W-7405-ENG-36.

2. Leveraging Comparative Sequencing Information to Generate a Complete Functional Map of Human Chromosome 19

Lisa Stubbs, Xiaochen Lu, Sha Hammond, Eddie Wehri, Anne Bergmann, Robin Deis, Angela Kolhoff, and Joomyeong Kim
Genomics Division, Lawrence Livermore National Laboratory
stubbs5@llnl.gov

As part of the Joint Genome Institute comparative sequencing team, we recently reported initial analysis of comparative alignments between the draft sequence of human chromosome 19 (HSA19) and related regions in the mouse. These initial studies identified a large number of conserved sequence elements, totaling ~5% of HSA19 DNA, which represent a rich source of candidate genes and exons in addition to the promoters, enhancers and regulatory sequences that control their tissue-specific expression.

We are presently focusing on carrying HSA19 annotation to the next stage, by providing a complete catalog of functionally verified genes and other biologically active sequence elements along the length of this small and extraordinarily gene-dense human chromosome. To do this, we have focused on tying expressed sequences together to define the full set of HSA19 and related mouse genes, confirming predicted genes and defining their 5'- and 3' borders. To gain clues to the biological functions of each gene, we are also determining cell-type-specific expression patterns systematically, through in situ hybridization of sectioned mouse and human tissues. We are also testing candidate promoters and

enhancers for function using a high throughput reporter assay system in cultured cells. These studies are designed to leverage DOE's long-term investment in HSA19 sequencing, provide a publicly accessible guide to the function of all 1200 genes and the location of associated regulatory sequences throughout the chromosome. These collected data will also permit us to associate regulatory sequence structure with function in both species, providing an unprecedented look at the composition and evolution of promoters, enhancers and other regulatory sequences and their evolution in mammals.

3. The Finishing of Human Chromosomes 19 and 5

Jane Grimwood, Jeremy Schmutz, Mark Dickson, Richard M. Myers, and all members of the Sequencing Group at the Stanford Human Genome Center
The Stanford Human Genome Center and the Department of Genetics, Stanford University School of Medicine, Palo Alto, CA 94304
jane@shgc.stanford.edu

For the last two years, the Stanford Human Genome Center has been collaborating with the Joint Genome Institute to generate high-quality finished sequence from "draft" sequences produced by the JGI. To date, we have submitted 176 Mb of finished sequence with an estimated error rate of 1 in 342,000 basepairs. This current collaboration continues through December 2002, by which time we will have finished both human chromosomes 19 and 5. We will discuss the current status of the chromosomes and the procedures we are using to obtain closure.

Chromosome 19, estimated to be around 58 Mb in length, comprises slightly more than 2% of the human genome. It is extremely gene rich, containing perhaps twice the number of genes per DNA sequence length than the rest of the human genome. It is also extremely repetitive, has a very high GC content and a skewed distribution of CpG islands. For these reasons, the finishing of this chromosome has been, and continues to be, a great challenge. Currently, 61% of the sequence is in a finished form and the remainder is in active finishing. Many small sequence gaps exist between clones and these are being filled by walking directly from spanning clones. The finishing of five highly repetitive areas of the chromosome is being attempted by breaking the repeats down into smaller cosmid units to isolate copies of the repeats.

Chromosome 5 is estimated to be 184 Mb in length. Currently, 77 Mb of the chromosome is in finished sequence form. An additional 50 Mb of the chromosome is in active finishing, with the remainder of the clones being brought up to full draft coverage by the Joint Genome Institute. Mapping efforts are continuing at the JGI to obtain complete clone coverage of this chromosome.

4. Assembly and Analysis of Finished Sequence for Human Chromosome 19

Anne Olsen[1], Susan Lucas[1] and the JGI Production Sequencing Group; Jane Grimwood[2], Jeremy Schmutz[2] and the Stanford Finishing Group; Laurie Gordon[3] and the LLNL Mapping Group; Paramvir Dehal[1], Art Kobayashi[1], Sam Pitluck[1] and the JGI Informatics Group; and Trevor Hawkins[1].
[1]DOE Joint Genome Institute, Walnut Creek, CA
[2]Stanford Human Genome Center, Palo Alto, CA
[3]Lawrence Livermore National Laboratory, Livermore, CA
olsen2@llnl.gov

Chromosome 19 has an estimated size of ~65 Mb and is the most GC-rich human chromosome. It also stands out as the chromosome with the highest content, relative to size, of repetitive sequences, CpG islands, and genes. The BAC/cosmid map of ch19 constructed at LLNL consists of seven contigs spanning ~98% of the estimated 58 Mb comprising the p- and q-arms. Map coverage extends to within 25 kb of the p-telomere (Riethman, http://www.wistar.upenn.edu/Riethman/) and into subtelomeric repeats on the q-arm. The most proximal several hundred kb on both the p- and q-arms have a high content of alpha satellite sequence, indicating proximity to the centromere. Mapping effort continues to close the few remaining map gaps, with TAR cloning (Kouprina et al.) in progress for four gaps that have been resistant to closure by other methods. A tiling path of clones spanning the chromosome has been sequenced, with 43.3 Mb (75% of the p- and q-arms) currently in a finished state. Finished sequence assembles into 154 contigs

with an average contig size of 280 kb. Analysis of the sequence indicates over 1200 known genes, including a large number of clustered gene families, as well as several hundred predicted genes. The distribution of repeats differs markedly from that of the genome as a whole, with chromosome 19 exhibiting a much higher density of Alu repeats and lower content of LINE sequences than the genomic average. A comparison of genetic and physical distances across the chromosome indicates several regions of sex-specific enhanced recombination, with an especially high male recombination rate towards the telomeres. Comparative studies with mouse and Fugu (Dehal et al.) are providing further insights into the genomic organization and evolution of this chromosome.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, the Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098, and the Los Alamos National Laboratory under contract No. W-7405-ENG-36.

5. Finishing of Human Chromosome 16

Norman Doggett, Mark Mundt, David Bruce, Cliff Han, Levy Ulanovsky, Larry Deaven, Susan Lucas, Trevor Hawkins, and JGI Staff
DOE Joint Genome Institute and Center for Human Genome Studies, Los Alamos National Laboratory
doggett@lanl.gov

Finishing of human chromosome 16 is being coordinated and conducted by the JGI's Los Alamos Center for Human Genome Studies. Coordination, and progress toward finishing is being monitored with a minimal tiling path clone map providing current active status for each clone and gap. The minimal tiling path map consists of 738 clones, 91 of which are cosmids and the remainder predominately BACs. These provide close to complete coverage of the 89 Mb of euchromatin. There are 17 clone gaps which are being closed by a combination of BAC end sequencing analysis and library screening. Most of the draft sequence for the chromosome was generated by the JGI with some draft contributions from WIBR (44 BACs) and WUGSC (26 BACs). As of November 1, 2001, the centers which have contributed significantly toward the finished sequence of clones in the tiling path include LANL (19.8 Mb), SHGC (12.3 Mb), TIGR (9.9 MB), and SC (1.5 Mb). The unique total of the chromosome completed as of November 1, 2001 is approximately 40 Mb. LANL will finish most of the remaining 49 Mb and we anticipate that a total of 50-60 Mb of the chromosome will have been completed by the time of this DOE conference and that the full euchromatin arms will be completed by this spring. We are finishing by managing the whole chromosome at once and integrating sequencing strategies, robotics, and an information management system into a highly automated process (see abstracts by Mundt et al. and Bruce et al.). Current status of chromosome finishing will be presented.

Supported by the US DOE, OBER under contract W-7405-ENG-36.

6. An Overview of the Finish Sequencing Process at LANL: Design, Automation, and Organization

David C. Bruce, Mark O. Mundt, Levy E. Ulanovsky, Heather A. Blumer, Judy M. Buckingham, Connie S. Campbell, Mary L. Campbell, Olga Chertkov, J. Joe Fawcett, Valentina M. Leyba, Kim K. McMurry, Linda J. Meincke, A. Christine Munk, Beverly A. Parson-Quintana, Donna L. Robinson, Elizabeth H. Saunders, Judith G. Tesmer, Linda S. Thompson, Patti L. Wills, Norman A. Doggett, and Larry L. Deaven
DOE Joint Genome Institute and Center for Human Genome Studies, Los Alamos National Laboratory
dbruce@lanl.gov

The challenge of high-throughput finishing is being addressed at Los Alamos National Laboratory (LANL) by integrating sequencing strategies, information management system, automation and personnel organization. The personnel are organized into specialized teams; Informatics, Subclone Re-Array, Template Preparation, Template Labeling, Oligonucleotide Synthesis, DENS, Sequencing,

Subcloning, End Sequencing, and Gap Closure. All samples are handled in a 96 or 384 well format. Library re-array is done using a Genetix Q-Bot or Packard MultiProbe robots. Template purification using a solid-phase reversible immobilization (SPRI) method features Robbins Hydra and TiterTek MultiDrop automations. Thermal cycling is done in 384 well format using MJ Research Tetrads. Primer synthesis is done in 96 well format using Mermaid oligonucleotide synthesizers (See abstract of Thompson, et al) or differential extension with nucleotide subsets (DENS, see abstract of Ulanovsky, et al). Labeling reaction strategies include Big Dye terminator, Big Dye primer, and Big Dye dGTP terminator chemistries. Labeled template is run on capillary ABI PRISM 3700 DNA Analyzers in a 384 well format. The teams are coordinated by instructions generated by an information management system (see abstract of Mundt, et al.).

Supported by the US DOE, OBER under contract W-7405-ENG-36.

Sequencing Resources

7. Construction of BAC Libraries Using Sheared DNA

Kazutoyo Osoegawa, Chung Li Shu, and Pieter J. de Jong
Children's Hospital Oakland Research Institute, Oakland, CA 94609
kosoegawa@mail.cho.org

Bacterial artificial chromosome (BAC) libraries have initially been developed to provide intermediate DNA substrates for genome mapping and sequencing. After completion of the human draft sequence, mapped and sequenced BAC clones have also become important for disease diagnostics and functional genomics. There is nevertheless still a need for additional BAC clones for regions poorly represented in the "conventional" BAC libraries to complete genome projects and to create more representative libraries for future genome projects. To this end, we cloned sheared DNA in a modified BAC vector in anticipation of reduced cloning bias. BAC libraries with different average insert sizes and random ends support a hybrid approach to genome sequencing based on a combination of whole genome shotgun and clone-by-clone sequencing. High-molecular-weight DNA is sheared by multiple cycles of freezing and thawing. The fragment ends are then blunted by treatment with Mung Bean nuclease and T4 DNA polymerase, and are ligated to the blunt-end side of an adapter which has a 3' overhang (ACAC) at the other end. The ligation products are size-fractionated to remove the excess of adapter and to obtain the desirable-size insert DNA fragments for cloning. The new vector (pTARBAC6) has two BstXI restriction sites flanking a replaceable stuffer fragment. Upon BstXI digestion, a vector fragment with two 3' overhangs (GTGT) is generated, complementary to the adapter-ligated genomic fragments. We have been able to construct several BAC libraries from *Drosophila*, *Ciona savignyi* and mouse with different average insert sizes to fit the applications. Provisional results with the new libraries indicate a random clone distribution and a very low level of undesirable chimeric clones. Initial screening results for the fly BAC library indicates a possible extension of contigs towards the telomeres. To facilitate closing the clone gaps in the human genome, we are constructing a BAC library using sheared DNA. Information on our completed libraries can be found at: www.chori.org/bacpac.

The US DOE specifically funded the technology development of sheared BAC libraries and the construction of a human BAC library (ER62962).

8. BAC Library End Sequencing in Support of Whole Genome Assemblies

David C. Bruce, Mark O. Mundt, Kim K. McMurry, Linda J. Meincke, Donna L. Robinson, Norman A. Doggett, and Larry L. Deaven
DOE Joint Genome Institute and Center for Human Genome Studies, Los Alamos National Laboratory
dbruce@lanl.gov

The Center for Human Genome Studies at Los Alamos National Laboratory has end sequenced over 75,000 BAC clones from *Fugu*, *Ciona*, *Chlamydomonas* and Human libraries to support whole genome shotgun sequencing and assembly efforts by the Joint Genome Institute of *Fugu*, *Ciona*, and *Chlamydomonas* and in support of our human chromosome 16 finishing efforts. Beginning from libraries arrayed in 384 well plates, stock plates are translated into a 96 well format. After growth in a 96 well deep plates, the sequencing template is purified using 96 well LigoChem ProPrep BAC 96 kits. Following sequencing template resuspension, the template is labeled with ABI PRISM BigDye Terminator v3.0 chemistry in 384 well format. The labeled template is run on ABI PRISM 3700 DNA Analyzer. We are achieving an overall 80% paired end pass rate and greater than 450 bp read length. Process details and quality statistics will be presented.

Supported by the US DOE, OBER under contract W-7405-ENG-36.

9. An Approach to Filling Gaps in the Sequence of the Human Genome

X.-N. Chen[1], P. Bhattacharyya[1], S. Y. Zhao[2], M. Sekhon[3], J. McPherson[3], M. Wang[4], U.-J. Kim[4], H. Shizuya[4], M. Simon[4], and J. R. Korenberg[1]
[1]Medical Genetics, Cedars Sinai Medical Center, UCLA, Los Angeles, CA
[2]The Institute for Genomic Research, Rockville, MD
[3]Washington University Genome Sequencing Center, St Louis, MI
[4]Caltech, Pasadena, CA
Xiao-Ning.Chen@cshs.org

The story of metazoan evolution is a story of genomic duplication. Primates are not an exception and the human genome reflects a rich history of recent duplication events that are a source of contemporary genomic variability and instability. We now link these duplicated regions to the draft sequence (Golden Path and Celera) and show that they are located throughout chromosome arms, reflect regions of instability and represent gaps in the current draft sequence of the human genome. To avoid biases in sequence sets introduced by unstable regions, we have defined at random a subset of BACs for putatively duplicated regions and integrated them with the draft sequence. They provide anchor points for sequencing centromeres, pericentromeres and duplications in chromosome arms. These include a total of 6,000 BACs mapped by FISH, 3,500 defined at random, 184 from screens with alpha satellite, 346 with telomeric oligos and ~2,000 from other screens of the Caltech BAC libraries A and B. About 957 are STS linked. Out of 6,000 BACs, 373 mapped to centromeric regions, 192 to single centromeres, 150 to multiples and 20 to all human centromeres. Of 990 multisite BACs, 350 were defined at random suggesting a minimum of 10% of the genome was duplicated and interspersed.

Fingerprint database analysis:
A total of 489 were fingerprinted, 33 with 5-29 bands showed no database match and suggested a minimum of 8% of duplications (non centromeric) were not represented in the fingerprint database.

End sequence analysis:
Golden Path 1.1 draft sequence analysis: Of the 434 end sequenced BACs, 134 or 30% had no match in the draft sequence; 145 had hits of over 98% homology and 147 had hits of 80-98%. Three were located on orphan contigs.

Celera database analysis: Out of 1020 ends 243 represent BACs with a single end sequenced and 382 with both ends sequenced. Out of the 243 single BAC ends, 53% had no significant hits (defined by $\leq$ 97% homology); 47% had hits of $\geq$ 98% homology. Of the 382 BACs with both ends sequenced only 134 pairs of ends had hits on the same chromosome. Only 65 out of these were spaced in the correct range (BES within 80-300 Kb and $\geq$ 350bp match to draft). Perhaps, the most important observation was that 14% (86 of 625 BACs) had no matches to the Celera database. Therefore, they identify the holes in the current human draft sequence.

This analysis of both sources of Human draft sequence (Golden Path 1.1) and Celera database suggests that at least 65% of BACs recognizing more than one site in the Human genome identified largely at random by FISH, were not included in the draft sequence and therefore identify gaps in both sources of the genome draft sequence. These BACs provide anchors for defining hotspots of genomic instability, for sequencing centromeric regions containing genes and for filling gaps in the draft sequence.

10. Isolation of Segments Missing from the Draft Human Genome Sequence Using Yeast

N. Kouprina, G. Solomon, S.-H. Leem, A. Ly, E. Pak, J. C. Barrett, and V. Larionov
Laboratory of Biosystems and Cancer, National Cancer Institute, NIH, Bethesda, MD 20892
kouprinn@mail.nih.gov

The reported draft human genome sequence includes multiple short contigs (groups of overlapping segments) that are separated by gaps of unknown sequence. The gaps in the draft sequence may arise from chromosomal regions that are not present in the *Escherichia coli* libraries used for DNA sequencing because they can not be cloned efficiently, if at all, in bacteria. To estimate the extent of the human genome missing in *E. coli* libraries, we compared euchromatic human DNA cloned in YACs and BACs. To isolate human genomic sequences in

yeast, we applied the Transformation-Associated Recombination (TAR) cloning method. This method allows selective cloning in yeast without DNA manipulations in vitro and avoids chimeric recombinants. The TAR cloning vector contained both YAC and BAC cassettes that allowed propagation of the same sequence in yeast and bacteria. Approximately 6% of human DNA sequences transformed less efficiently and was less stable in *E. coli* than in yeast. This fraction included both specific genes (KAI1 and MUC2) and anonymous DNA regions that have not previously been recovered from BAC libraries. DNA sequences from the ends of these YAC clones are not in the draft genome sequence. The results suggest that it may be possible to fill gaps in the draft human sequence using clones propagated as YACs in yeast. We demonstrate the use of recombinational cloning in yeast (TAR) to recover problematic genomic regions and to verify contigs assembly rapidly and potentially systematically.

11. Recent Segmental Duplications: A Dynamic Source of Gene Innovation and Complex Regions of Sequence Assembly

J. A. Bailey, J. E. Horvath, M. E. Johnson, M. Rocchi, and **E. E. Eichler**
Department of Genetics and Center for Human Genetics, Case Western Reserve School of Medicine and University Hospitals of Cleveland, Cleveland, OH, 44106
eee@po.cwru.edu

It has been estimated that 5% of the human genome consists of interspersed duplicated material that has arisen over the last 30 million years of evolution. Two categories of recent duplicated segments can be distinguished: segmental duplications between non-homologous chromosomes (transchromosomal duplications) and duplications largely restricted to a particular chromosome (chromosome-specific duplications). A large proportion of these duplications exhibit an extraordinarily high degree of sequence identity at the nucleotide level (>95%) spanning large (1-100 kb) genomic distances. Through processes of paralogous recombination, these same regions are targets for rapid evolutionary turnover among the genomes of closely related primates. The dynamic nature of these regions in terms of recurrent chromosomal structural rearrangement and their ability to generate to create fusion genes from juxtaposed cassettes suggests that duplicative transposition has been an important force in the evolution of our genome. Cycles of segmental duplication over periods of evolutionary time may provide the underlying mechanism for domain accretion and the increased modular complexity of the vertebrate proteome. Further, our data suggest that a small fraction of important human genes may have emerged recently through duplication processes and will not possess definitive orthologues in the genomes of model organisms. I will discuss computational methods developed in my laboratory to 1) unambiguously identify recent genomic duplicates within the human genome and 2) to assess their importance in hominoid gene innovation. The impact of this chromosomal architecture for assembly the final draft sequence, particularly within chromosomes 16 and 19, will be discussed.

12. Pooling DNA Clones for Shotgun Sequencing

Richard Gibbs[1] and the staff of the Baylor College of Medicine-Human Genome Sequencing Center, Wei Wen Cai[2], and Allan Bradley[3]
[1]Baylor College of Medicine-Human Genome Sequencing Center
[2]Department of Molecular and Human Genetics, Baylor College of Medicine
[3]Sanger Center
agibbs@bcm.tmc.edu

We have developed two methods based upon clone pooling for more efficient shotgun DNA sequencing. The first is Concatenation cDNA Sequencing (CCS), a procedure where multiple cDNA inserts are joined together by ligation for sequencing in a single shotgun project. CCS has been continually refined since our first experiments with small numbers of pooled cDNAs in 1995. In the month of August 2001 we completed 900 cDNAs using the method. With further increments in the efficiency of the approach we expect to have the ability to analyze entire mammalian transcriptomes in a few months.

The second methodology is an improvement on procedures for the sequencing and assembly of whole genomes. The Clone Array Pooled Sequencing Scheme (CAPSS), is based upon the pooling of rows and columns of arrayed genomic clones prior to shotgun library construction. Random sequences are accumulated, and the data processed by sequential comparison of rows and columns, to assemble the sequence of clones at points of intersection. Compared to either a clone-by-clone approach or whole genome shotgun sequencing, CAPSS requires relatively few library constructions and only minimal computational power for a complete genome assembly. Computer simulations show the practicability of the method and testing of CAPSS in the assembly of Rat Genome sequences is underway.

13. Production Clone Rearraying Using the QBot (Genetix Ltd.) and the LANL Cherrypicking Program

John J. Fawcett, James Colehan, Lyn Honeyborne, Bill Stevenson, David C. Bruce, Norman A. Doggett, and Larry L. Deaven
DOE Joint Genome Institute and Center for Human Genome Studies, Los Alamos National Laboratory and Genetix Limited, United Kingdom
fawcett@lanl.gov

Clone rearray or cherry picking of subclones is the first hands on step toward setting up finishing reactions. Finishing plates contain up to 96 unique candidate subclones selected from thousands of archival source plates. Cherrypicking is directed rearraying of clones from source plates into one or more destination plates. We have automated this process on the Qbot with custom software. The LANL Cherrypicking program utilizes QBot capabilities accessible via the Developer's Toolkit (Genetix Ltd). Subclones from source plates are deposited into specific destination wells and plates as specified by imported finishing scripts. Rearrayed subclone plates are provided to DENS and custom primer finishing teams for appropriate finishing reactions (see Abstract of Bruce et al.). The development of LANL Cherrypicking program was a joint effort of LANL and Genetix Ltd.

14. Applications of Isothermal Rolling Circle Amplification in a High-Throughput Sequencing Environment

John C. Detter, Jamie M. Jett, Andre R. Arellano, Alicia R. Ferguson, Kristie Tacey, Mei Wang, Heidi C. Turner, Susan M. Lucas, Ken Frankel, Paul Predki, Dan Rokhsar, Paul M. Richardson, and Trevor L. Hawkins.
U.S. DOE Joint Genome Institute, Walnut Creek, CA 94598
detter2@llnl.gov

High-throughput sequencing requires several DNA amplification steps. In general, researchers have been limited to methods such as *in vivo* amplification in *E. coli.* and Polymerase Chain Reaction to obtain source DNA for library creation and template DNA for sequencing. Replication by rolling circle is common among bacteriophages and viruses in nature. Recently, Rolling Circle Amplification (RCA) with Φ29 DNA polymerase has been applied *in vitro* to specific target sequences using specific primers and to circular cloning vectors using random hexamer primers to achieve exponential DNA amplification by way of DNA strand displacement. At the Joint Genome Institute we have examined the use of random hexamer primed RCA (TempliPhi™) for several applications related to sequencing. Here, we demonstrate that RCA can be used effectively for amplification of plasmids, cosmids and BACs for direct end sequencing. DNA from RCA amplified BAC and Cosmid clones can also be used to generate random shotgun libraries. In addition, we show that whole bacterial genomes can be effectively amplified from cells or small amounts of purified genomic DNA without apparent bias for use in downstream applications including whole genome shotgun sequencing.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, the Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098, and the Los Alamos National Laboratory under contract No. W-7405-ENG-36.

15. Efficient Isothermal Amplification of Single DNA Molecules

Stanley Tabor and Charles Richardson
Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115
tabor@hms.harvard.edu

We are developing DNA polymerases for use in DNA sequencing and amplification applications. We will describe a very efficient isothermal amplification system that we have been developing that is based on the replication machinery of bacteriophage T7. It is capable of amplifying a single DNA molecule, increasing the amount of DNA up to a trillion-fold in a 30 min reaction. Amplification is nonspecific. The template can be circular (e.g. plasmid or BAC DNA) or linear (e.g. genomic DNA). The products are linear double-stranded DNA fragments that average several thousand base pairs in length. The reaction requires the T7 DNA polymerase , the T7 helicase/primase complex (T7 gene 4 protein) , and single-stranded DNA binding protein. The reaction requires no exogenous primers, using the inherent primase activity of the T7 primase. It is critical to remove all contaminating DNA from the reaction mixture, since the system efficiently amplifies all DNA present. We have developed a successful strategy to deal with this problem that involves cleansing the reaction mixture by treatment with Micrococcal nuclease.

There are a number of applications in which a robust generic amplification system is attractive:

1. A simple alternative to the current methods used to prepare plasmid and BAC DNA templates for DNA sequencing.
2. The immortalize of rare genomic DNAs, such as hard-to-culture microorganisms or purified chromosomes.
3. An extremely sensitive assay for the presence of DNA in a sample. By including a dye in the reaction mixture that fluoresces when it binds DNA, the reaction provides a fast, robust assay that detects DNA over 13 orders of magnitude in several minutes.
4. The sequencing of haplotypes can be expedited by the use of templates that have been amplified from single chromosomes.

We will present our progress on the use of isothermal amplification in these applications.

16. Amplification of BAC DNA with Rolling Circular Amplification

Cliff S. Han, Judy Tesmer, Linda L. Meincke, Donna L. Robinson, Connie S. Campbell, Larry L. Deaven, and Norman A. Doggett
DOE Joint Genome Institute and Center for Human Genome Studies, Los Alamos National Laboratory
han_cliff@lanl.gov

Bacteria artificial chromosome (BAC) cloning systems provide the major clone resource in sequencing the human genome and will continue to be one of the major tools in finishing the human genome and for sequencing other organisms. Despite the many advantages of BACs, it remains a challenge to purify large amounts of BAC DNA in a high throughput manner because of its low copy number and relatively long DNA strands. Here we introduce a method based on rolling circular amplification to generate large amount of BAC DNA. Starting with less them 1 ml culture, several micrograms of BAC DNA can be generated. The amplified DNA is well suited for restriction mapping, BAC end sequencing, and subcloning for shotgun sequencing. The major steps of the protocol include 1) lysis of bacteria cell with lysozyme to release DNA; 2) degrading of bacteria DNA with restriction enzyme and plasmid-safe DNAse; 3) amplification of BAC DNA with on rolling circular amplification.

Supported by the US DOE, OBER under contract W-7405-ENG-36.

17. A Single-Copy, Amplifiable Plasmid Vector That Uses Homing Endonuclease Recognition Sites to Facilitate Bidirectional Nested Deletion Sequencing of Difficult Regions

John J. Dunn, Laura Praissman, Laura-Li Butler-Loffredo, and Sean McCorkle
Biology Department, Brookhaven National Laboratory, Upton, NY 11973-5000
jdunn@bnl.gov

The long term goal of this project is to develop improved methods for finishing difficult regions in draft sequences. Difficult regions we are focusing on include long repeats and regions that interfere with polymerase progression. Towards this goal, we have developed a plasmid vector, pSCANS, based on the low-copy F replicon which allows rapid generation of an ordered set of nested deletions from either strand of a cloned DNA fragment. The size of the vector has been reduced to the 4.4-kbp range by removing the 2.5-kbp sop (stability of plasmid genes) region from the F replicon. The resulting plasmid has the low copy number typical of F plasmids and it remains stable enough to be easily maintained by growth in the presence of kanamycin, the selective antibiotic. DNA in amounts convenient for sequencing is readily obtained by amplification from an IPTG-inducible P1 lytic replicon. The vector's multiple cloning region (MCR) has several unique sites for both shotgun and directional cloning. It is flanked on one side by recognition sequences for the extremely rare cutting intron encoded nucleases I-CeuI and I-SceI, and on the other side by a recognition sequence for another intron encoded enzyme, PI-PspI and a nicking site for the phage f1 protein, gpII, that initiates f1 rolling circle DNA replication. Cleavage with the intron encoded enzymes leaves four-base 3' overhangs that are resistant to digestion with *E. coli* ExoIII. Between these sites and the MCR are recognition sites for several rare 8-base cutters that leave ExoIII sensitive termini. Double cutting with one intron encoded enzyme and an adjacent rare cutting restriction endonuclease allows for unidirectional 3' to 5' digestion across the insert with ExoIII.

Alternatively, plasmid linearized on one side of an insert with I-SceI can be blunt ended to produce an ExoIII sensitive end and then cut with I-CeuI to generate an adjacent ExoIII resistant end.

The f1 nicking site can be used for ExoIII digestion of the other strand of the insert or for producing single-stranded plasmid circles for library normalization or subtraction. After ExoIII digestion, the resulting single-stranded regions are digested with S1 nuclease, and the ends are repaired and ligated with T4 DNA polymerase and ligase. Pooling samples from several different Exo III digestion time points before subsequent S1 treatment generates a good distribution of deletion clones following electroporation. Deletion clones are sized and sequenced using vector specific forward and reverse primers.

Cloned fragments at least 10 thousand base pairs long can be sequenced and assembled easily by generating an ordered set of nested deletions whose ends are separated by less than the length of sequence read from a single priming site within the adjacent vector. Assembly of the overlapping sequences is guided by knowledge of the relative length of the portion of the fragment remaining in the clone, as determined by gel electrophoresis. Even highly repeated DNA can be assembled correctly at comparatively low redundancy by knowing the relative locations of the sequences obtained.

Nested deletions can also demarcate the ends of "problem regions" that obstruct polymerase progression which causes a failure of the sequencing reaction. Several different approaches can then be used to attempt to finish the sequence. One promising method is to PCR amplify the "problem region" and completely replace guanine with 7-deaza guanine. Incorporation of this analogue prevents formation of non-Watson-Crick base paired DNA triplexes that otherwise block some sequencing reactions. The amplicons are then sequenced using standard Big dye terminator chemistry supplemented with various reagents know to facilitate extension through problematic regions. Our results with a C+T–rich repeat from chromosome 19 will be presented.

18. DENS: Finishing Without Custom Primers

Levy Ulanovsky, Olga Chertkov, Malinda Stalvey, Marie-Claude Krawczyk, David Hill, David Bruce, Mark Mundt, Larry Deaven, and Norman Doggett
DOE Joint Genome Institute and Center for Human Genome Studies, Los Alamos National Laboratory
levy@lanl.gov

DENS (Differential Extension with Nucleotide Subsets) is primer walk sequencing without custom primer synthesis. DENS largely eliminates the cost of custom primer synthesis - several dollars, compared to less than a dollar for the rest of the expenses (per lane) combined. DENS works by converting a short primer (selected from a pre-synthesized library of 1440 octamers with 2 degenerate bases each) into a longer one on the template at the intended site only. DENS starts with a limited initial extension of the octamer primer at 20° C in the presence of only 2 of the 4 possible dNTPs. The primer is extended by 5 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 cycle-sequencing at 60° C accepts the primer extended at the intended site, but not at alternative sites where the initial extension (if any) is generally short. We have now automated all labor-intensive steps in DENS and have employed this as part of our finishing strategy to improve low quality targets. Several megabases of chromosome 16 have been finished using > 40,000 DENS reactions with the success rate rising from ~ 40% to ~ 80%.

Supported by the US DOE, OBER under contract W-7405-ENG-36.

19. High Throughput Synthesis of Oligonucleotides in Support of Finishing

L. Sue Thompson, Mark Mundt, David Bruce, Larry Deaven, and Norman Doggett
DOE Joint Genome Institute and Center for Human Genome Studies, Los Alamos National Laboratory
thompson_1_sue@lanl.gov

Los Alamos is currently using a Liquid Chemical Dispensing Robot, built by Bioautomation and called the MerMade by its creators at the University of Texas Southwest, to synthesize large numbers of oligonucleotides for use in custom primer finishing reactions. With the first MerMade installed in February of 1999, approximately 12,000 oligos were made the first year. During and since that time methods and infrastructure have been modified and developed to optimize protocols for cost effective and safe production of large numbers of oligonucleotides. One important determination made during the first year of operations was that a more effective ventilation system was needed to minimize hazardous chemical exposure to the technician and surrounding laboratory tenants. The Mermade oligonucleotide synthesis laboratory completed a move in August of 2000 to the site of a chemical synthesis laboratory facility with two custom-design fume hoods with individual Phoenix controls for the existing Mermade and a second MerMade. Each MerMade synthesizer is designed to synthesize two standard 96 well plates of oligonucleotides in a single run, using standard phosphoramidite chemistry. Synthesizing four days per week on two MerMades allows the production of 1536 oligonucleotides per week and 70,000 oligonucleotides per year. A Beckman Biomek 2000, automated workstation, performs most time consuming multi-channel pipetting tasks. With the Normalization Wizard software each plate is quantitated on a Molecular Dynamics SpectraMax plate reader and normalized to the user's specifications. Quality control involves running a representative sample of each plate on a gel and/or analyzing a representative sample on the Voyager DE Biospectrometry Workstation with MALDI-TOF mass analysis.

Supported by the US DOE, OBER under contract W-7405-ENG-36.

20. Automated 384-Well Purification for Terminator Sequencing Products

Chris Elkin, Hitesh Kapur, David Humphries, Troy Smith, and Trevor Hawkins
U.S. DOE Joint Genome Institute, Walnut Creek, CA 94598
Elkin1@llnl.gov

We have developed an automated purification method for terminator sequencing products based on magnetic bead technology. This four-step method is optimized for use in 384-well PCR plates and low costs. The end product is essentially salt free and allows for water loading onto capillary gel systems. We have tested this method with various DNA templates such as PCR, Plasmids, Cosmids and Rolling Circle Amplification products and found a 40 base pair increase in read length, as compared to ethanol precipitation methods. Our new method also eliminates all centrifugation steps and is compatible with both MegaBACE 1000 and ABI Prism 3700 instruments. Currently, this method is producing 100 (384-well plates) per day on a Biomek FX robotic platform with an average pass rate of 90% and readlength > 600 (Q20) bp.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, the Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098, and the Los Alamos National Laboratory under contract No. W-7405-ENG-36.

21. Whole Genome Direct Sequencing: Completion of Microbial Genome and Mammalian BAC Projects using ThermoFidelase, Fimer and D-Strap Technologies

S. Kozyavkin, A. Malykh, K. Mezhevaya, A. Morocho, N. Polouchine, V. Shakhova, O. Shcherbimina, and A. Slesarev
Fidelity Systems, Inc., 7961 Cessna Avenue, Gaithersburg, MD 20879-4117
serg@fidelitysystems.com
http://www.fidelitysystems.com

We have developed a novel strategy for genomic DNA sequencing that minimizes the number of reactions and potentially eliminates the need in subcloning and production of shotgun libraries. The successful scale up of our approach has resulted in complete and highly accurate sequence of a microbial genome and a number of human and mouse BACs.

A core component of the procedure is the use of genomic DNA as a template in a robust sequencing reaction. The addition of ThermoFidelase 2 with its unique combination of topoisomerase and DNA binding activities is used to shorten the cycles of denaturation and primer annealing. The dramatic increase in specificity, quality and yield of priming from megatemplates is achieved by using Fimers (modified oligonucleotides with proprietary SUC modifications) instead of regular primers and multiplying the number of thermal cycles. The third element of new strategy, D-Strap is based on Fimer design that targets evolutionary conserved elements in RNA- or protein-coding genes. We have optimized reagents and protocols for a sequencing production environment of a small team and limited resources.

Using a novel approach, we have determined the complete 1,694,969 nucleotide sequence of the GC-rich genome of *Methanopyrus kandleri*, a hyperthermophile that can grow at 110 deg C. As little as 3.3x sequencing redundancy was sufficient to assemble the genome with < 1 error per 40 kb. The optimization of protocols and reagents resulted in the increase of an average read length of direct genomic traces from 370 q20 bases in the beginning of the project to 500 bases at the end. Due to the unique position of *M. kandleri* on the phylogenetic tree (a single species phylum in euryarchaeal division), the initiation step based on D-Strap was supplemented with limited sequencing of cloned plasmids directly from cell cultures (i.e., without isolation of DNA). The utility of produced Fimers was further demonstrated in sequencing reactions with the other strain of *M. kandleri* (9% sequence difference). We continue the development of D-Strap technology for direct sequencing of microbial genomic DNA of various sizes and taxonomic origin (~ 5 Mb, 20% sequence difference).

The completeness and high quality of *M. kandleri* sequence was a prerequisite for the application of COG-based methods in comprehensive genome annotation, analysis of proteome evolution and reconstruction of cellular metabolism which was done by Dr. Koonin's group at NCBI in a very short period of time (weeks). We anticipate that combination of low redundancy direct genomic sequencing and speedy analysis will help eliminate backlog of unfinished projects and make microbial and comparative genomics more affordable for small scientific teams.

Whole Genome Direct sequencing of mammalian organisms (3 Gb genomes) can not be done with the current technology. Instead, BAC libraries with sequenced ends and low coverage Whole Genome Shotgun (WGS) data can be used to initiate the project. We have optimized Fimer design and BAC sequencing protocols for the production of reads of up to ~ 1 kb long, including sequencing through difficult and repetitive regions. The utility of D-Strap Fimers that target evolutionary conserved mammalian exons was demonstrated on human and mouse BACs. Our data show that 100% contiguity and high quality of assembled sequence can be achieved starting from <3x WGS data and producing a low number of direct reads off human or mouse BAC templates. The critical elements for the robust genome finishing technology and methods for further optimization of overall workflow for high throughput environment will be discussed.

Supported in part by DOE and NIH (DE-FG02-98ER82577, 00ER83009, R44GM55485, R43HG02186).

22. A Tape Conveyer System for Storage and Distribution of Biological Samples

Ger van den Engh and Juno Choe
Institute for Systems Biology, Seattle, WA 98115
engh@systemsbiology.org

We are developing a tape system for packaging large numbers of biological samples. The samples are stored in 5 microliter wells that are formed in a long plastic tape. A cover tape seals the wells. A 10 inch diameter spool can hold 10,000 samples. The tapes can be used for storage and retrieval of cells, microorganisms, or biological molecules. When used as conveyer system, the tapes can be used to perform experiments on large numbers of samples.

The tapes are particularly powerful when combined with cell sorting. A cell sorter may deposit a string of rare event on a tape. The content of each well may be expanded by the PCR or by natural proliferation single cells.

One application is the rapid subcloning of a piece of DNA. DNA fragments are transfected into a specially constructed plasmid. The plasmid has a cloning site in between green and red fluorescent protein. The native plasmid transcribes a hybrid protein that fluoresces red when excited with blue/green light. When the linker between the proteins is disrupted by an insert, the transcribed protein emits green fluorescence. The bacteria carrying plasmids with an insert have a different color from the bacteria that do not have an insert. Thus the bacteria with inserts can be easily detected in a cell sorter. The use of sorting for subclone selection represents a significant increase in speed in clone preparation for DNA fragment sequencing.

23. Developing a High Throughput Lox Based Recombinatorial Cloning System

Robert Siegel[1], Raj Jain[2], Nileena Velappan[2], Leslie Chasteen[2], and **Andrew Bradbury**[2]
[1]Pacific Northwest National Laboratory, Richland, Washington
[2]Los Alamos National Laboratory, Los Alamos, New Mexico
amb@lanl.gov

The selection of antibodies (single chain Fvs – scFvs) against protein targets can be done using a number of different systems, including phage, phagemid, bacterial or yeast display vectors. Genetic selection methods have also been developed based on yeast two hybrid and enzyme complementation systems. In general, selection vectors are not suitable for subsequent scFv production. Furthermore, once scFvs have been selected, they can be usefully modified by cloning into other destination vectors

(e.g. by adding dimerization domains, detection domains, eukaryotic expression in eukaryotic vectors etc.). However, this is relatively time consuming, and requires checking of each individual construct after cloning. An alternative to cloning involves the use of recombination signals to shuttle scFvs from one vector to another. These have the advantage that DNA restriction and purification can be avoided. Such systems have been commercialized in two general systems: Gateway™, uses lambda att based recombination signals, while Echo™ uses a single lox based system to integrate a source plasmid completely into a host plasmid.

We have examined the potential for using heterologous lox sites and cre recombinase for this purpose. Five apparently heterologous lox sites (wild type, 511, 2372, 5171 and fas) have been described. A GFP/lacZ based assay to determine which of these were able to recombine with each other was designed and implemented. Of the five, three (2372, 511 and wt) were identified which recombined with one another at levels less than 2%.

To use recombination as a cloning system, it is important to be able to select against host vectors which do not contain the insert of interest. Two toxic genes were examined for this purpose. The tetracycline gene confers sensitivity to nickel, while the sacB gene confers sensitivity to sucrose. We confirmed these sensitivities, although found that some antibiotic resistances interfere with survival of bacteria hosting non-tetracycline containing plasmids.

In preliminary experiments we have demonstrated that recombination from one plasmid to another, using 2272 and wild type lox sites and sacB or tetracycline, can occur in vivo at very high efficiency. This opens the possibility of using this system to easily transfer scFvs after selection to other plasmids. However, the utility of this system is not limited to scFvs – any DNA fragment (gene, open reading frame, promoter etc.) can easily be shuttled from one plasmid to another using these lox based signals.

24. Plant Mini-Chromosome Vectors

J. Mach and H. Zieler
Chromatin, Inc.
mach@chromatininc.com

Chromatin's technology focuses on the design of plant mini-chromosomes, large DNA molecules with the capacity to carry multiple genes. Other gene delivery methods for plants introduce individual genes into a host chromosome, causing irreparable damage to host genes and unpredictable effects on the expression of the introduced gene. In contrast, because plant mini-chromosomes segregate independently from the host chromosomes, they eliminate insertional mutagenesis and position effects. A major obstacle to the development of mini-chromosome has been the challenge of identifying centromeres. Chromatin's proprietary technology allows purification of centromere DNA from important crop species and incorporation of that DNA into mini-chromosomes, along with other essential chromosomal components. These mini-chromosomes can be delivered into plant cells and tested to determine which sequence combinations exhibit the highest degree of stability through successive cell generations. Like the mini-chromosomes developed previously for yeast and bacterial systems, plant mini-chromosomes will improve the reliability and efficiency of gene delivery, enable precise control of gene expression, and significantly expedite the analysis of new gene functions.

Chromatin's technology focuses on the design of plant mini-chromosomes, large DNA molecules with the capacity to carry multiple genes. Other gene delivery methods for plants introduce individual genes into a host chromosome, causing irreparable damage to host genes and unpredictable effects on the expression of the introduced gene. In contrast, because plant mini-chromosomes segregate independently from the host chromosomes, they eliminate insertional mutagenesis and position effects. We are currently assembling key chromosomal components into plant mini-chromosomes. These mini-chromosomes can be delivered into plant cells and tested to determine which sequence combinations exhibit the highest degree of stability through successive cell generations. Like the mini-chromosomes developed previously for yeast and bacterial systems, plant

mini-chromosomes will improve the reliability and efficiency of gene delivery, enable precise control of gene expression, and significantly expedite the analysis of new gene functions.

25. Sampling Diversity with Mitochondrial Genomics

Jeffrey L. Boore, Nikoletta Danos, David DeGusta, H. Matthew Fourcade, Lisa Gershwin, Allen Haim, Kevin Helfenbein, Martin Jaekel, Kirsten Lindstrom, J. Robert Macey, Susan Masta, Mónica Medina, Rachel Mueller, Marco Passamonti, Corrie Saux, Renfu Shao, and Yvonne Vallès
DOE Joint Genome Institute, Walnut Creek, CA 94598
JLBoore@lbl.gov

Mitochondrial DNA (mtDNA) comparisons serve as a model for genome evolution and as a tool for reconstructing evolutionary relationships. Relative to the nuclear genome, this system has several advantages for a comprehensive sampling across life. Mitochondrial genomes are small and gene-rich with a conserved complement of genes that are homologous among plants, protists, fungi, and animals. The products of these genes participate in well-characterized biochemical processes and play important roles in metabolism, health, aging, and biochemical adaptation. The comparison of mitochondrial genomes, especially of the relative arrangements of their genes, has proven to be among the best of datasets for reconstructing the evolutionary relationships among major groups of organisms. MtDNAs are typically circular, allowing physical isolation from nuclear DNA. To date, individual labs have produced, at most, a few mtDNA sequences per year. We are developing techniques to greatly accelerate this effort, including protocols for the rapid purification of mtDNAs for shotgun cloning, bioinformatics tools for streamlined data processing, web-based tools for comparisons of mitochondrial genomic features, and improved computational methods for reconstructing minimum genome rearrangement pathways. These innovations can lead to a larger, more comprehensive survey of biodiversity at the genome level than has been previously imagined possible.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, the Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098, and the Los Alamos National Laboratory under contract No. W-7405-ENG-36.

Instrumentation

26. Method for Fast and Highly Parallel Single Molecule DNA Sequencing

Jonas Korlach, Michael Levene, Stephen W. Turner, Harold G. Craighead, and Watt W. Webb
School of Applied and Engineering Physics, Cornell University, Ithaca, NY 14853
jk109@cornell.edu

There is an urgent need for the ability to sequence single molecules of DNA with long read lengths to provide intrinsic haplotyping and SNP profiling capability, enable the detection of extremely rare species or strains in a sample, and eliminate prior amplification steps.

Our project of developing such a DNA sequencing method is based on the proposition that the temporal order of base additions during DNA polymerization of a single nucleic acid molecule can be measured in real time. The activities of single molecules of DNA polymerase are observed in a detection format allowing the separate observation of many individual molecules simultaneously, thus creating a very fast and highly parallel new method of DNA sequence acquisition. The properties of the enzyme and the format of data acquisition facilitate evolution of this technology into compact and inexpensive DNA sequence analysis systems, suitable for rapid *de novo* or resequencing of DNA using unprocessed, unpurified samples. Hundreds of sequencing reactions could be performed simultaneously, with read lengths of individual reactions ranging into the tens to hundreds of kilobase pair size range.

In the course of developing this technology, challenges had to be solved regarding (i) the design, synthesis and evaluation of suitable fluorescently labeled nucleotide analogs which are both amenable to single molecule detection and efficiently utilized by the polymerase, and (ii) the fabrication of nanostructured devices permitting detection of single molecule polymerase activity at high fluorophore concentrations. Single base incorporation events were observed recently using this technology, proving the principal validity of this novel sequencing approach.

As efficient DNA synthesis occurs only at substrate concentrations much higher than the pico- or nanomolar regime typically required for single molecule analysis, zero-mode waveguide nanostructures are described as a way to overcome this limitation. They effectively reduce the observation volume to tens of zeptoliters (10^{-20} l), thereby enabling an increase in the upper concentration limit amenable to single fluorophore detection. Zero-mode waveguides thus extend the range of biochemical reactions that can be studied on a single molecule level into the micromolar range.

Supported by DOE grant DE-FG02-99ER62809.

27. Fast Detection of Nucleic Acid Hybridization with a Tapered Optical Fiber Sensor

Hyunmin Yi[2], Vildana Hodzic[2], James J. Sumner[2], Matthew P. Delisa[2], Saheed Pilevar[2], Frank H. Portugal[2], James B. Gillespie[2], Christopher C. Davis[2], and **William E. Bentley**[1]
[1]Center for Agricultural Biotechnology and Department of Chemical Engineering, University of Maryland, College Park, MD 20742
[2]University of Maryland, College Park, MD 20742
bentley@eng.umd.edu

A tapered single-mode optical fiber sensor was used to investigate gene regulation by hybridization of nucleic acids. The sensor is based on the evanescent field excitation of fluorescence of surface-bound fluorophores. The same tapered optical fiber is used to excite and collect fluorescence. This sensor design can potentially eliminate a number of problems typically encountered in other systems. Use of infrared fluorophores prevents background signals from natural visible region fluorescence, making extensive purification procedures and (RT)PCR amplification unnecessary. A pulsed mode operation prevents photobleaching of the dye and enables multiple detections. Further, the sensor surface is

easily regenerated and the signal detection is nearly instantaneous for on-line monitoring of the target gene transcription without additional analyte labeling. Various derivatives of fluorescein were used to quantitate chemically modified sensor surfaces and test alternative chemical crosslinkers. Oligonucleotides of 20 base pairs with functional groups at either end (5' or 3') were covalently / noncovalently incorporated onto the sensor surface and used as probe molecules for complementary analyte strands in the samples. Regulation of *dnaK* gene in a high cell density culture of *Escherichia coli* was investigated as a model system. Finally, Near-field Scanning Optical Microscopy (NSOM) and evanescent wave excitation of surface-bound fluorophores on a prism surface were used to further investigate surface behavior.

28. High Performance Capillary Electrophoresis in DNA Sequencing and Analysis: Recent Developments

Barry L. Karger, Lev Kotler, Arthur Miller, and Hui He
Barnett Institute, Northeastern University, 360 Huntington Avenue, Boston, MA 02115
b.karger@neu.edu

Our laboratory has devoted efforts to enhance separation of Sanger sequencing fragments to increase productivity of DNA sequencing. In the past, we developed linear polyacrylamide (LPA) to be the matrix with the longest read lengths (1300 bases in 2 hours). Recently, we focused in detail on the reasons for this performance. We compared N,N-dimethylacrymide with LPA and found that a key advantage of LPA is its ability to achieve high performance at 70-75° C, whereas the more hydrophobic polymer significantly loses separation above 50° C, due to a less robust entanglement structure. A second factor in our long read length ability relates to a base-caller that is able to read sequence down to peak resolution as low as 0.25. We have also recently shown that read lengths of 975 bases can be achieved in 40 min. by careful attention to the denaturants used in the buffer. These results will be described.

Additionally, we will describe the use of an automated fraction collector for capillary array electrophoresis for determination of mutant species. Mutants are separated from wild type by constant denaturant capillary electrophoresis. The collected components are then sequenced for determination of the specific mutations. The automated multiple array fraction collection approach can be also used for determination of differentially expressed cDNA's, especially useful for organisms whose genomes are not yet sequenced. We will illustrate these applications in this poster.

29. Microchannel DNA Sequencing by End-Labeled Free Solution Electrophoresis (ELFSE): Development of Polymeric End-Labels, Wall Coatings, and Electrophoresis Methods

Wyatt N. Vreeland, Jong-In Won, Robert J. Meagher, M. Felicia Bogdan, and Annelise E. Barron
Northwestern University, Dept of Chemical Engineering, Evanston IL USA 60208
w-vreeland@northwestern.edu

High-performance DNA separation matrices required for sequencing, which are based on viscous, entangled polymer solutions, require application of high pressure for rapid loading into the narrow-diameter channels used in current state-of-the-art microchannel electrophoresis sequencing instruments. We are working to develop a new method of *free-solution* DNA sequencing based on the separation of DNA-protein bioconjugates, which still uses the Sanger ddNTP chain termination reaction but obviates the need for a "gel". As such, this new method, called *End-Labeled Free Solution Electrophoresis* (ELFSE) is especially amenable to high-field electrophoresis in microfluidic chips.

ELFSE functions through the attachment of a monodisperse, uncharged, polymeric "molecular drag-tag" to the 5' terminus of DNA sequencing fragments. The fixed amount of hydrodynamic drag this frictional label engenders for each Sanger sequencing fragment enables sized-based separation of DNA by electrophoresis in free solution (*i.e.*, in the absence of a sieving matrix). The scaling law for

the electrophoretic mobility of a sequencing fragment is given by equation 1:

$$\mu = \frac{\rho(N-\beta)}{\xi(N+\alpha)} \qquad (1)$$

where ρ and ξ are the charge and effective hydrodynamic friction of a single DNA base respectively, N is the number of DNA bases in the sequencing fragment, and β and α are the amount of charge and drag of the end label in units of DNA bases. Ideally, $\beta = 0$, that is, the labels are uncharged. Thus to enable long-read DNA sequencing, polymeric end-labels that engender substantial hydrodynamic drag are required, as labels that provide low amounts of hydrodynamic drag are only effective in separating DNA fragments of short lengths.

The production of a high molar mass, uncharged *and* completely monodisperse polymer that will not substantially interact with microchannel walls and that has a unique point for attachment of DNA sequencing fragments is not a trivial task. Specifically our efforts have focused on determining what chemical characteristics are necessary for an optimal drag-tag, and development of cloning methodology to allow the production of high-molar mass protein polymers. To this end, we have developed a unique cloning strategy to produce protein polymer end-labels of large and controlled length (up to 1000 amino acids) from synthetic genes in *E.coli*. To date we have employed ELFSE-type separations for the molar mass profiling of the end-label molecules, highly multiplexed genotyping of clinically relevant DNA samples *via* Single Base Extension (SBE) methodology, and demonstrated separation of end-labeled DNA sequencing fragments. We have developed a highly hydrophilic, adsorptive polymer wall coating that enables high-resolution separation of DNA-protein conjugates. Current challenges we are addressing include optimization of the chemical nature of the protein polymer end-label as well as of the electrophoretic separation techniques and protocols.

30. Microfabricated Fluidic Devices for the Analysis of Genomic Materials

K. A. Swinney, R. S. Foote, C. T. Culbertson, S. C. Jacobson, and **J. Michael Ramsey**
Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831-6142
ramseyjm@ornl.gov

We are developing monolithic microfabricated fluidic devices for the analysis of genomic materials including DNA, peptides, proteins and cells. Microfluidics offers many potential advantages for performing automated and rapid analyses of small quantities of biological materials. Most strategies for analysis of genomic materials include a chemical separation process. Microfabricated separation devices have typically provided a separation performance equivalent to conventional laboratory technology using orders of magnitude less sample materials and taking one-to-two orders of magnitude less time.

One of the limitations of microfabricated separation devices is the resolving power achievable with in a small footprint, e.g., a few centimeters on a side. We have been investigating strategies that could allow greater separative performance using short separation distances. Such capabilities are necessary to keep devices small enhancing practicability and reducing costs. Microfabricated devices and biochemical strategies that allow the comprehensive analysis of proteins are also being developed. Comprehensive two-dimensional separations have been demonstrated for peptides that yield a peak capacity of approximately 1000. Improvements of a factor of at least four appear possible. We are attempting to move this technology toward the analysis of protein mixtures. Moreover the use of microfluidics for the automated analysis of the contents of single cells is being pursued.

The goal is to automate the loading of cells with reagents, incubation, lysis and analysis of cellular contents through chemical separations. Initial results have been obtained from a device that accomplishes the latter two steps. Jurkat cells have been loaded with Oregon Green and the automated analysis of individual cells produces an electropherogram for

each one showing the Oregon Green and its metabolites. The analysis rate in this case was approximately 10 cells/min. Various aspects of these devices will be described.

31. Molecular Gates for Improved Sample Cleanup and Handling in Microfabricated Devices

Tzu-Chi Kuo, Donald M. Cannon, Mark A. Shannon, Paul W. Bohn, and **Jonathan V. Sweedler**
Department of Chemistry, Department of Mechanical Engineering, and Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, 600 South Mathews Avenue, Urbana, Illinois 61801
sweedler@scs.uiuc.edu

The development of integrated systems capable of automated accurate sequence generation from sample introduction to sequence output is an important goal of the DOE Human Genome Project. Microfabricated DNA analyzers (such as microfabricated PCR systems with integrated CE systems) have been developed that offer a number of important advantages compared to the traditional large-scale methods. However, the actual interface between these microfabricated subassemblies can be problematic.

Extension of microfabricated microfluidic devices to three-dimensions opens new vistas for applications and parallels the massively three-dimensional architectures characteristic of electronic devices. Externally controllable interconnects, employing nuclear track-etched polycarbonate membranes with nanometer diameter pores, are described that produce hybrid three-dimensional fluidic architectures. Using nanofluidic structures to connect microfluidic channels allows a variety of flow control concepts to be implemented, leading to hybrid fluidic architectures of considerable power and versatility. The key distinguishing characteristic feature of nanofluidic channels is that fluid flow occurs in structures of the same size as physical parameters that govern the flow. Furthermore, the separations capacity factor, k', governed by the surface-to-volume ratio, can be quite large. For example, k' increases by ~120 when a 200 nm i.d. nanopore with a 10 nm thick coating is compared with a 20 micron i.d. wall-coated open tubular column with the same coating. These nanofluidic interconnects can be thought of as fluidic diodes, albeit with a much richer array of parameters to control biasing.

Forward/reverse bias is controlled by applied potential, surface charge density (pH controllable), ionic strength, and even by the characteristics of the fluidic network in which the interconnect is placed. We demonstrate the use of these interconnects to collect an analyte band from an electrophoretic separation and transport the band to another fluidic layer. The construction and operating characteristics of these devices is described for a variety of applications.

The successful molecular gate has the ability to transfer a particular DNA band on-device after PCR analysis to another fluidic layer to allow sample cleanup and even to capture a particular DNA band eluting from the separation channel for further characterization. This allows easier interfacing between the separate components of a total "lab-on-a-chip" sequencer. We have optimized- molecular gate technology and are developing the protocols for high efficiency and fast sample capture and release.

32. Electron Tomography of Whole Cells

Grant J. Jensen and **Kenneth H. Downing**
Lawrence Berkeley National Laboratory
khdowning@lbl.gov

Recent advances suggest that whole microbial cells could be imaged by electron tomography to "molecular" resolution, sufficient to locate and identify large macromolecular complexes in their native state within their cellular contexts. Such an advance could prove crucial for the success of the Microbial Cell Project, as no other existing imaging modality can be expected to provide the high resolution structural information necessary to characterize many gene products, track complex formation and pathways, finely localize structures and functions within the cell, and understand the detailed affects of future attempts to customize microbes. Electron tomography proceeds by quick freezing whole cells in a thin, aqueous film across an electron microscope grid, recording projection images of them from multiple directions in the

microscope, and combining these images in a computer to produce a three-dimensional reconstruction. We propose to develop and test this technique on two microbes: one chosen for its ideal imaging characteristics, and the other for its potential Department of Energy mission relevance.

This project combines the expertise of key pioneers in the fields of high resolution protein and cellular imaging by electron microscopy, an ideally equipped electron microscope, and a uniquely well suited model system that will clearly reveal the potential contribution of electron tomography towards the goals of the Microbial Cell Project.

33. Cast Thy Proteins Upon the Water: Fluid Proteomics in a 2–D World

Barry Moore, Chad Nelson, Mike Giddings, Mark Holmes, Melissa Kimball, Norma Wills, John Atkins, and Ray Gesteland
Department of Human Genetics, University of Utah, Salt Lake City, Utah, USA
barry.moore@genetics.utah.edu

Genome sequencing has been an exceptionally successful endeavor with completion of over 60 non-viral genomes, and another 573 underway. Understanding the functional products of these genomes – proteins – has developed into the perhaps more difficult, but nonetheless rapidly developing field of proteomics. Much of the energy and resources in proteome projects has been invested in peptide mapping by 2-D gel electrophoresis followed by MALDI-MS analysis. One shortcoming of the 2-D gel approach, however, is that it does not allow for determination of the full length protein mass as the proteins must be proteolyitcally digested in order to be removed from the gels.

We have developed a two dimensional LC-MS approach to protein identification. This approach allows us to determine the mass of the full length protein as well as peptide mapping of it's fragments – potentially providing information on post-translational modifications, signal peptide cleavage sites, cotranslational recoding events, and errors in ORF prediction found in the database. Furthermore, while most peptide mapping search algorithms rely on searching predefined ORFs, we have developed a searching algorithm that searches total genomic sequence for peptide matches thus allowing peptide mapping to identify programmed frameshifting events and short or ambiguous ORF proteins rejected by ORF prediction software, and thus unavailable for search by ORF based peptide mapping search algorithms.

We have applied these techniques to yeast mitochondria, and have identified a number of proteins, providing new information on the mass of predicted, but unknown proteins and corrections to database errors.

34. Single Cell Proteome Analysis— Ultrasensitive Protein Analysis of *Deinococcus radiodurans*

Shen Hu, Amy Dambrowitz, Roger Huynh, and **Norm Dovichi**
Department of Chemistry, University of Washington
dovichi@chem.washington.edu

We are developing technology to monitor changes protein expression in single tetrads of *D. radiodurans* following exposure to ionizing radiation. We hypothesize that exposure to ionizing radiation will create a distribution in the amount of genomic damage and that protein expression will reflect the extent of radiation damage. To test these hypotheses, we will develop the following technologies:

- Fluorescent markers for radiation exposure
- DNA/rRNA determination of each cell in a *D. radiodurans* tetrad
- Two-dimensional capillary electrophoresis analysis of the protein content of a single tetrad
- Ultrasensitive laser-induced fluorescence detection of proteins separated by capillary electrophoresis

These technologies will be combined to determine protein expression in single tetrads of *D. radiodurans*, the extent of DNA damage following exposure to Cs-137 radiation, and the amount of chromosomal and rRNA per cell. This technology will be a powerful tool for functional analysis of the

microbial proteome and its response to ionizing radiation.

35. High-Throughput SNP Scoring with GAMMArrays: Genomic Analysis Using Multiplexed Microsphere Arrays

P. Scott White, Hong Cai, David Torney, Lance Green, Diane Wood, Francisco Uribe-Romeo, LaVerne Gallegos, Julie Meyne, Paul Jackson, Paul Keim, and John Nolan
Bioscience Division and Theoretical Division, Los Alamos National Laboratory and Department of Microbiology, Northern Arizona University
scott_white@lanl.gov

We have developed a platform for the discovery and scoring of SNPs that is capable of meeting greatly increasing demands for high throughput and low cost assays. Called GAMMArrays, or Genomic Analysis using Multiplexed Microsphere Arrays, the basic platform consists of fluorescently labeled DNA fragments bound to microspheres, which are analyzed using flow cytometry. The platform provides no-wash assays that can be analyzed in less than 1 minute per sample, with sensitivities far superior to other approaches. SNP scoring is performed using minisequencing primers and fluorescently labeled dideoxynucleotide terminators. Furthermore, by using commercially available sets of multiplexed microspheres it is possible to score dozens to hundreds of SNPs simultaneously. Multiplexing, when coupled with the high throughput rates possible with this platform makes it possible to score several million SNPs per day at costs that are a fraction of competing technologies.

GAMMArrays are enhanced by the use of universal oligonucleotide tags. These tags consist of carefully designed, unique DNA tails, or capture tags, incorporated into each minisequencing primer, that are complementary to an address tag attached to a discrete population of microspheres in a multiplexed set. This enables simultaneous minisequencing of large numbers of SNPs in solution, followed by capture onto the appropriate microsphere for multiplexed analysis by flow cytometry.

We present results from multiplexed SNP analyses of bacterial pathogens, and human mitochondrial DNA and HLA genetic variation. These analyses are performed on a small number of relatively large PCR amplicons, each containing numerous SNPs that are scored simultaneously. In addition, these assays are easily integrated into conventional liquid handling automation systems, and require no unique instrumentation for setup and analysis. Very high signal-to-noise ratios, ease of setup, flexibility in format and scale, as well as low cost of these assays make them highly versatile and extremely valuable tools for a wide variety of studies where SNP scoring is needed.

36. Characterization of the *D. radiodurans* Proteome using Accurate Mass Tags

R. D. Smith[1], G. A. Anderson[1], M. S. Lipton[1], L. Pasa-Tolic[1], J. Fredrickson[1], J. R. Battista[2], M. J. Daly[3], C. Masselon[1], R. J. Moore[1], M. F. Romine[1], Y. Shen[1], and H. R. Udseth[1]
[1]Pacific Northwest National Laboratory
[2]Louisana State University
[3]Uniformed Services University of the Health Sciences
rd_smith@pnl.gov

In our approach to microbial proteomics our objective is to circumvent the limitations of conventional approaches by directly characterizing the cell's polypeptide constituents using a combination of high resolution separations and the mass accuracy and sensitivity obtainable with Fourier transform ion cyclotron resonance (FTICR) mass spectrometry. Protein identification is based upon global approaches for protein digestion and accurate peptide mass analysis for the generation of "Accurate Mass Tags" (AMTs). Our two-stage strategy exploits FTICR for validation and subsequent routine measurement of peptide AMTs from "potential mass tags" initially identified using tandem mass spectrometry methods, and thus providing the basis for high throughput proteome-wide measurements. A single high resolution capillary liquid chromatography separation combined with high sensitivity, high resolution and accurate FTICR measurements has been shown to be capable of characterizing peptide mixtures of more than 100,000 components, sufficient for broad

protein identification in microbial systems. Attractions of the approach include the capability for automated high-confidence protein identification, broad and unbiased proteome coverage, and the capability for exploiting stable-isotope labeling methods for quantitative relative protein abundance measurements. Using this strategy, we have been able to identify AMTs for >60% of the potentially expressed proteins in the organism *Deinococcus radiodurans*. Approximately 32% and 16% of the ORFs from the *D. radiodurans* database are predicted to be hypothetical (having no significant homology to any proteins in any other public genome sequence databases at the time of annotation) and conserved hypothetical (having limited homology to a functionally uncharacterized ORF), respectively. We identified 48% of these hypothetical proteins and 55% of the conserved hypothetical proteins. The approach will also be shown to allow the detection of modified proteins, as well as quantitative measurements of changes in protein abundances resulting from environmental perturbations.

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

37. Combining "Top-Down" and "Bottom-Up" Mass Spectrometry Approaches for Proteomic Analysis: *Shewanella oneidensis* — A Case Study

Robert Hettich, Nathan VerBerkmoes, Jonathan Bundy, James Stephenson, Loren Hauser, and Frank Larimer
Oak Ridge National Laboratory
hettichrl@ornl.gov

The current rapid expansion of the field of proteomics will help this become one of the key components of the DOE-BER Genomes-To-Life Project, in particular assisting in the determination of gene function for the Microbial Genomics Program. The development of methods for rapid, large-scale mass spectrometry (MS) analyses of proteins from complex biological samples is considered to be critical for proteome studies. Two major approaches for MS-based proteome analysis have been employed to date. In the most common "bottom-up" approach, proteins are separated, proteolytically digested, and subsequently identified via MS analysis of the resultant peptides. An alternative approach, termed the "top-down" method, involves analysis and identification of intact proteins via accurate mass measurement. Since an intact protein mass is measured, this method may be advantageous for the detection of post-translational modifications, which may be missed in analyses by the bottom-up approach, where only a fraction of the total peptide population may be detected.

We have developed a novel method for proteome analysis that integrates both the top-down and bottom up approaches, capitalizing on the unique capabilities of each method. Bacterial cellular lysates were initially fractionated via anion exchange liquid chromatography. Each fraction, which consisted of intact protein species, was divided into two samples. The first sample of each fraction was digested with trypsin protease and analyzed with reversed phase-LC-MS/MS for protein identification. This two dimensional separation strategy greatly enhanced the number of proteins that could be identified as compared to a similar analysis of an unfractionated lysate. The second sample of each fraction was analyzed on an electrospray Fourier transform mass spectrometer for high-resolution identification of the intact proteins with the top-down approach. The use of the two methods in concert enabled the facile detection of such common post-translational modifications as loss of N-terminal methionine and signal peptide cleavages. This new approach was applied in a preliminary proteomic analysis of *Shewanella oneidensis*, a metal reducing microbe of potential importance to DOE in the field of bioremediation. With this experimental MS approach, it was possible to identify over 700 proteins from *S. oneidensis*, with the identification including ribosomal proteins, hypothetical proteins, suspected metal reducing proteins, and membrane proteins. One key protein, an ATP-binding protein, was found a have a molecular mass that was about 3 kDa lighter than that expected from the gene sequence. Detailed inspection of high-resolution MS data indicated that

a 20 amino acid signal peptide had been cleavaged off the N-terminus of the protein. This experimental data is being used to refine bioinformatic tools that are used to predict the presence and identity of signal peptides. We feel that this level of information is a unique capability of our combination "top-down" — "bottom-up" MS proteomic method.

38. Oligonucleotide Mixture Analysis via Electrospray and Ion/Ion Reactions

Scott A. McLuckey[1], **Jin Wu**[1], Jonathan L. Bundy[2], James L. Stephenson, Jr.[2], and Gregory B. Hurst[2]
[1]Department of Chemistry, Purdue University, West Lafayette, IN 47907-1393
[2]Chemical and Analytical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6365
mcluckey@purdue.edu, jwu@purdue.edu

Electrospray ionization combined with ion/ion reactions in a quadrupole ion trap can be used for the direct analysis of oligonucleotide mixtures. Elements to the success of this approach include factors related to ionization, ion/ion reactions, and mass analysis. This work deals with issues regarding the ion polarity combination, viz., positive oligonucleotides/negative charge transfer agent versus negative oligonucleotides/positive charge transfer agent. Anions derived from perfluorocarbons appear to be directly applicable to mixtures of positive ions derived from electrospray of oligonucleotides, in directly analogy with positive protein ions. Conditions for forming positive oligonucleotide ions devoid of adducts were more difficult to establish than for forming relatively clean negative oligonucleotide ions. A new approach for manipulating negative ion charge states in the ion trap is described and is based on use of the electric field of the positive charge transfer agent for storage of high mass negative ions formed during the ion/ion reaction period. Oxygen cations are shown to be acceptable for charge state manipulation of mixed-base oligomers but induce fragmentation in poly-adenylate homopolymers. Protonated isobutylene ($C_4H_9^+$), on the other hand, is shown to induce significantly less fragmentation of poly-adenylate homopolymers. This presentation will describe the results mentioned above along with new developments in instrumentation that will lead to improvements in mass analysis. Two new systems are expected to come on line in the coming months. One is comprised of a three dimensional quadrupole ion trap with at least twice the mass-to-charge range of the previous system and two to four times the resolving power. The other system will employ trapping in a two-dimensional quadrupole ion trap followed by time-of-flight mass analysis. This system promises to provide improvements in mass resolution, mass accuracy, and speed.

39. Peptide Sequencing and Identification Using de novo Analysis of Tandem Mass Spectra

William R. Cannon, K. D. Jarman, and K. H. Jarman
Pacific Northwest National Laboratory
william.cannon@pnl.gov

Algorithms for sequencing peptides from mass spectrometry data have been seen as increasingly important as researchers start to think beyond genomic data towards the study of proteins. In many proteomic analyses, tandem mass spectrometry is used to identify peptides, which can then be linked back to their parent protein. Presently, most analyses of proteomic tandem mass spectra rely on comparisons with static information within genomic sequence databases. For example, the peptide dissociation pattern from MS/MS is compared to hypothetical dissociation patterns for peptides that are computationally generated from a sequence database. As such, the ability to discover new knowledge about post-translational modifications, mutations, unexpected reading frames and sequencing errors is severely compromised.

An alternative is to derive the protein sequence information de novo by exploiting the sequence information contained in tandem mass spectra. That is, since a MS/MS experiment on a biopolymer results in a sequential fragmentation of the parent biopolymer into daughter fragments, it is possible to determine the sequence of the parent peak from only the mass spectrum of the daughter peaks and knowledge of the mass of each monomeric subunit of the polymer. Here we report on the use of graph theory and set covering techniques that incorporate

peak intensities as well as mass values to identify peptides without reliance on sequence databases.

40. Laser Desorption Mass Spectrometer for DNA Sequencing and Hybridization Detection

Winston C. H. Chen[1], Steve L. Allman[1], Klara J. Matteson[2], and Lauri Sammartano[3]
[1]Oak Ridge National Laboratory
[2]University of Tennessee, Medical Center
[3]St. Olaf University
chenc@ornl.gov

During the past few years, we have developed various approaches to sequence DNA and to measure DNA hybridization with mass spectrometries.

For DNA sequencing, both Sanger's enzymatic synthetic method and Maxam Gilbert's chemical degradation method have been pursued. Single stranded DNA with size up to 130 nucleotides and double stranded DNA with size up to 200 base pairs have be sequenced with laser desorption mass spectrometry. One approach is to use matrix-assisted laser desorption/ionization (MALDI) for DNA detection. We also developed a novel approach with laser induced acoutic desorption for DNA detection. In addition to the sequencing with DNA ladders, we also developed direct DNA sequencing technology without the need of DNA ladders. This technology is particularly suitable for primer and DNA probe analysis.

Microarray DNA hybridization has been considered as a valuable tool for high throughput DNA analysis. We have developed mass spectrometry technology for DNA hybridization analysis. With mass spectrometry as a detector, multiplexing hybridization on a single hybridization spot can be achieved. Thus, the throughput can be further increased. Furthermore, it is easier to distinguish perfect hybridization from single base mismatched hybridization. SNP (single nucleotide polymorphism) can be quickly analyzed with this approach. Hybridization with genomic DNA with mass spectrometry detection has also been demonstrated. Experimental details will be presented in poster.

41. Novel Molecular Labeling for Post-Genomic Studies

Xian Chen, Tom Hunter, Fadi Abdi, Haining Zhu, John Engen, Songqing Pan, Sheng Gu, Li Yang, Morton Bradbury, and Vahid Majidi
C-ACS, Chemistry Division, Bioscience Division, Los Alamos National Laboratory
chen_xian@lanl.gov

Now that scientists have mapped the human genome, an even greater challenge has come to light: making sense of vast amount of information contained in a genome. This challenge can be broken down into two main areas: functional genomics, which involves the further and accurate study of DNA sequence diversity to understand their function, and proteomics, the study of the full repertoire of proteins encoded by a genome. Mass spectrometry (MS) is a promising tool for rapid, rigorous, and sensitive analyses in both areas, but critical advances are needed to increase its specificity and accuracy for the analyses at the genomic level. To address these cutting-edge issues, we have developed a novel MS-based Mass Tagging technique. On a genomic scale, our technique uses stable-isotope-labeled precursors—particular nucleotides for DNA, or amino acids for proteins—to label DNA or protein molecules residue-specifically for MS analysis. Displayed through a characteristic mass-split pattern induced by labeled precursors, the content of particular nucleotides or amino acid residues in particular DNA or protein fragments can be readily determined. We have applied this strategy successfully to many aspects of functional genomics and proteomics, including screening SNP, validating DNA sequencing data with ambiguities left-open by gel electrophoresis, identifying cellular proteins including membrane-bound and scarce proteins, detecting post-translational modifications in a residue-specific manner, and analyzing contact interfaces in protein/protein complexes. Our strategy of nucleotide- or amino acid-specific mass tagging in DNA or protein molecules provides a much more sensitive and accurate way of molecular labeling than radiological or chemical labeling. In addition to

the parameter of mass-to-charge ratio (m/z), the use of these site-specific stable-isotope labels tagging biological molecules in a sequence-specific way have dramatically enhanced the specificity, accuracy, sensitivity, and throughput of the MS-based technology for functional genomics and proteomics analyses.

Our publications in the past two years are included as follows

1. Xian Chen, Zhengdong Fei, Lloyd M. Smith, E. M. Bradbury, and Vahid Majidi (1999) Stable isotope-assisted MALDI-TOF mass spectrometry allows accurate determination of base compositions of PCR products. *Anal. Chem.* **71**:3118-3125.
2. Xian Chen, Lloyd M. Smith, and E. M. Bradbury (2000) Site-specific mass tagging with stable isotopes in proteins for accurate and efficient protein identification. *Anal. Chem* **72**:1134-1143.
3. Fadi Abdi, E. Morton Bradbury, Norman Doggett, and Xian Chen (2001) Rapid identification of single nucleotide polymorphism using mass-tagged deoxynucleotide. *Nuclei. Acid Res.* **29**:13 e61.
4. Tom Hunter, Li Yang, Haining Zhu, E. Morton Bradbury, Vahid Majidi, and Xian Chen (2001) Rapid identification of yeast complex mixtures using peptide mass mapping constrained with stable isotope-tagged peptides. *Anal. Chem.* **73**:4891-4902
5. Haining Zhu, Tom Hunter, Songqin Pan, E. Morton Bradbury, and Xian Chen (2001) Residue-specific Mass Signatures for the Efficient Identification of Protein Modifications by Mass Spectrometry. *Anal. Chem.* In press
6. Songqin Pan, E. Morton Bradbury, and Xian Chen (2001) Unambiguous sequencing of in vitro transcripts using MALDI-TOF MS coupled with stable isotope-enriched nucleotides. Submitted to *Nuclei. Acid Res.*
7. John R. Engen, E. Morton Bradbury, and Xian Chen (2001) Using stable-isotope labeled proteins for hydrogen exchange studies in complex mixtures. Submitted to *Anal. Chem.*

42. Monolithic Integrated PCR Reactor-CE Microsystem for DNA Amplification and Analysis to the Single Molecule Limit

Eric T. Lagally[1], Chung N. Liu[2], and Richard A. Mathies[1,3]
[1]UCB/UCSF Joint Bioengineering Graduate Group, Berkeley, CA 94720
[2]Department of Chemical Engineering, University of California, Berkeley, CA 94720
[3]Department of Chemistry, University of California, Berkeley, CA 94720
lagally@zinc.cchem.berkeley.edu

Microfabrication is an effective method for creating integrated microfluidic devices for high-performance chemical and biochemical analysis (1-3). Our early work included the development of the first integrated PCR reactor with a CE system (4) and the development of a CE chip with an integrated electrochemical detector (5). More recently, we have devised a monolithic system for conducting the polymerase chain reaction (PCR) directly connected to a capillary electrophoresis (CE) microchannel for product separation and analysis (6). Samples are loaded precisely into a 150-300 nL PCR reactor using 50 nL valves and hydrophobic vents. The sample is cycled between three temperatures using a resistive heater mounted on the bottom of the chip, and amplification products are directly injected and separated on the capillary electrophoresis channel. The device takes as little as 30 seconds/cycle, representing a vast improvement over conventional thermal cycling systems, which can take up to 5 minutes/cycle.

Our PCR-CE device has recently demonstrated successful detection of a PCR product amplified from a single DNA template molecule, bringing this technology to the limiting molecular sensitivity of the PCR reaction (7). In this work, a single M13 DNA template molecule was co-amplified with a control that was outside the stochastic regime. Calculation of peak area ratios between the two template products reveals discrete clusters, and these clusters conform to the expected Poisson distribution for stochastic single molecule events with a mean occupancy of 0.9 molecules. This device has the highest molecular sensitivity ever demonstrated in a PCR chip device. Previous static systems required

~6,000 starting copies (8), and continuous-flow geometries required as many as ~10^8 starting copies (9). High-sensitivity analyses, such as comparative gene expression studies from individual cells, can also be performed using these devices.

Recent improvements include the fabrication of a PCR-CE device with integrated resistance temperature detectors (RTDs) and heaters (10). Sputtered platinum four-wire RTDs are fabricated inside the PCR chambers and platinum heaters with gold leads are fabricated on the backside of the device using sputtering and electroplating processes. Both elements connect to outside electronics using standard PC board connectors. These integrated components have resulted in more accurate temperature measurement and more efficient heating of the PCR chamber than previously possible due to the closer proximity of the RTD to the sample and better thermal contact between the heater and the glass wafer. Successful amplification of a multiplex human sex-determination amplification of the centromeric alphoid repeat (11) in a one-step reaction from human buccal cells has recently been accomplished with this device.

We are also developing technology to enable a 96-sample PCR-CAE microplate for high-throughput integrated PCR-CE analyses from small amounts of template DNA. Current progress includes exploration of different valve and vent structures to reduce valve dead volumes and to increase the scalability and ease of fabrication of the device. Such an integrated device will also be capable of multiplexing with disparate amplification protocols while greatly reducing sample and reagent volumes and costs.

These results demonstrate a key advance in the development of an integrated microfluidic system that performs complete genetic analyses at sub-microliter volumes, useful in the areas of point-of-care genetics and rapid identification of infectious diseases.

References:

1. Liu, S., Shi, Y., Ja, W. W. Mathies, R. A. (1999) 71, 566-573.
2. Shi, Y., Simpson, P., Scherer, J. R., Wexler, D., Skibola, C., Smith, M. Mathies, R. A. (1999) Anal Chem 71, 5354-5361.
3. Simpson, P. C., Woolley, A. T. Mathies, R. A. (1998) Journal of Biomedical Microdevices 1, 7-26.
4. Woolley, A. T., Hadley, D., Landre, P., deMello, A. J., Mathies, R. A. Northrup, M. A. (1996) Anal. Chem. 68, 4081-4086.
5. Woolley, A. T., Lao, K., Glazer, A. N. Mathies, R. A. (1998) 70, 684-688.
6. Lagally, E. T., Simpson, P. C. Mathies, R. A. (2000) Sens. Actuator B-Chem. 63, 138-146.
7. Lagally, E. T., Medintz, I. Mathies, R. A. (2001) Anal. Chem. 73, 565-570.
8. Cheng, J., Shoffner, M. A., Hvichia, G. E., Kricka, L. J. Wilding, P. (1996) Nucleic Acids Res. 24, 380-385.
9. Kopp, M. U., de Mello, A. J. Manz, A. (1998) Science 280, 1046-1048.
10. Lagally, E. T., Emrich, C. A., Mathies, R. A. (2001) Lab Chip 1, in press.
11. Neeser, D. Liechtigallati, S. (1995) J. Forensic Sci. 40, 239-241.

43. Advances in Radial Capillary Array Electrophoresis Chip Sequencing and Genotyping Technology

Brian M. Paegel[1], Robert G. Blazej[2], Lorenzo Berti[1], Charles A. Emrich[3], James R. Scherer[1] and Richard A. Mathies[1]

[1]Department of Chemistry, University of California, Berkeley, CA 94720 USA

[2]UCB/UCSF Joint Bioengineering Graduate Group, University of California, Berkeley, CA 94720 USA

[3]Biophysics Graduate Group, University of California, Berkeley, CA 94720 USA

brian@zinc.cchem.berkeley.edu

In 1999, we introduced the radial microfabricated capillary array electrophoresis (μCAE) chip, rotary confocal fluorescence detection, and their application to simple genotyping (1). We have more recently been working on applying this technology to DNA sequencing with the goals of increasing analysis speed, integration and automation of sample preparation, and reduction in costly reagent consumption. μCAE devices can help us to accomplish these goals by allowing high-speed

DNA sequencing (2) and monolithic array construction for high-throughput analyses (3). Our current 96-lane μCAE device incorporates our radial array design (1) and hyper-turn channel geometries to extend separation channel lengths to 15.9 cm for DNA sequencing (4). Performing high-quality DNA sequencing required overcoming such challenges as high-viscosity gel matrix filling, system buffering for prolonged periods of electrophoresis, and sample and buffer evaporation at 60° C. Using a high-pressure gel-filling instrument (5) and the Berkeley radial confocal fluorescence scanner, we have demonstrated sequencing of 41,000 phred 20 bases of a M13mp18 sequencing standard in only 25 minutes (6). We are also working closely with the DOE's Joint Genome Institute to adapt the Berkeley μCAE platform for production scale sequencing samples.

We have also used the μCAE device to develop a new method for polymorphism identification and screening. Polymorphism ratio sequencing (PRS) employs a novel dye labeling and reaction scheme to produce sequencing extension ladders for unambiguous SNP scanning. Combined with μCAE, the entire human mitochondrial genome was compared to a reference sample in one run of a 96-lane sequencing device (30 min.). PRS leverages the exquisite sensitivity intrinsic to fluorescence techniques and internal controls to extend effective read length and improve accuracy. This is particularly advantageous in population studies to determine allelic frequencies. Titration studies on pooled DNA samples demonstrate minor allele frequency detection limits below 10% (7).

Development of the PRS method was facilitated by the production of novel energy-transfer (ET) dye-cassette labels. Cassettes contain a fluorescein donor moiety linked by a sugar-phosphate spacer to an emitter dye. The cassette terminates in a mixed disulfide, which can be easily coupled to thiol-activated oligonucleotides via disulfide exchange (8). Together, these technologies address some of the basic needs of current process biology efforts and will be the next-generation high-throughput DNA sequencing and genotyping methodology.

1. Shi, Y., Simpson, P. C., Scherer, J. R., Wexler, D., Skibola, C., Smith, M. T. and Mathies, R. A. (1999) *Anal. Chem.* **71,** 5354-5361.
2. Woolley, A. T. and Mathies, R. A. (1995) *Anal. Chem.* **67,** 3676-3680.
3. Simpson, P. C., Roach, D., Woolley, A. T., Thorsen, T., Johnston, R., Sensabaugh, G. F. and Mathies, R. A. (1998) *Proc. Natl. Acad. Sci. U. S. A.* **95,** 2256-2261.
4. Paegel, B. M., Hutt, L. D., Simpson, P. C. and Mathies, R. A. (2000) *Anal. Chem.* **72,** 3030-3037.
5. Scherer, J. R., Paegel, B. M., Wedemayer, G. J., Emrich, C. A., Lo, J., Medintz, I. L. and Mathies, R. A. (2001) *Biotechniques* **31,** 1150-1156.
6. Paegel, B. M., Emrich, C. A., Wedemayer, G. J., Scherer, J. R. and Mathies, R. A. (2001) *Proceedings of the National Academy of Sciences, U. S. A.*, (in press).
7. Blazej, R. G., Paegel, B. M., Emrich, C. A. and Mathies, R. A. (2001), (in preparation).
8. Berti, L., Medintz, I. L., Tom, J. and Mathies, R. A. (2001) *Bioconjugate Chem.* **12,** 493-500.

44. Integrated Platform for Detection of DNA Sequence Variants Using Capillary Array Temperature Gradient Electrophoresis

Zhaowei Liu[1], Cymbeline T. Culiat[2], Tim Wiltshire[3], Christina Maye[1], Heidi Monroe[1], Kevin Gutshall[1], and **Qingbo Li**[1]

[1]SpectruMedix Corporation, 2124 Old Gatesburg Road, State College, PA 16803

[2]Life Science Division, Oak Ridge National Laboratory, P.O. Box 2009, MS 8077, Oak Ridge, TN 37831

[3]Genomics Institute of the Novartis Research Foundation, San Diego, CA 92121

qbli@spectrumedix.com

With more sequence information available, studies of DNA sequence variants such as mutation and single nucleotide polymorphism (SNP) are an important next stage in genome research and disease studies. Although many techniques have been developed for the detection of sequence variants, sensitive, high-throughput, and flexible techniques are still required for accurate detection and characterization if large-scale scans for sequence variants are to be used effectively for establishing

correlations between certain variants and behavior of a biological system.

We have developed a highly versatile platform that performs temperature gradient capillary electrophoresis (TGCE) for mutation/SNP detection, sequencing and mutation/SNP genotyping for identification of sequence variants on an automated 24-, 96- or 192-capillary array instrument. In the first mode, multiple DNA samples consisting of homoduplexes and heteroduplexes are separated by capillary electrophoresis, during which a temperature gradient is applied that covers all possible Tms for the samples. The differences in Tms result in separation of homoduplexes from heteroduplexes, thereby identifying the presence of DNA variants. The sequencing mode is then used to determine the exact location of the mutation/SNPs in the DNA variants. The first two modes allow the rapid identification of variants from the screening of a large number of samples. Only the variants need to be sequenced. The third mode utilizes multiplexed single base extensions (SBEs) to survey mutations and SNPs at the known sites of DNA sequence. The TGCE approach combined with sequencing and SBE is fast and cost-effective for high throughput mutation/SNP detection.

We further test the capabilities and sensitivity of TGCE in mutation scanning in the mouse genome by scanning candidate genes for ENU induced mutations, and scanning a mutagenized ES cell library for specific gene mutations. To achieve this we used a test set of DNAs from a collection of 480 450-500bp PCR fragments that had previously been sequenced in 6 different mouse strains for SNP discovery so all mutations were known in the fragments. This set of DNAs contained a range of polymorphisms including single base pair changes, multiple SNPs and a few deletions and insertions as well (48% no SNP, 52% one or more SNPs). We used all 5x96 well samples as a test set, even those that gave no PCR product or more than one PCR product, and analyzed results for all data. The TGCE method implemented is a highly sensitive, high-speed, and high-throughput technique for mutation analysis. We report that 95% of the TGCE analysis concur with direct sequencing results. In addition, TGCE detected four more mutations that were initially missed by direct sequencing.

45. New Microfabrication Technologies for High-Performance Genetic Analysis Devices

Charles A. Emrich[2], Toshihiro Kamei[1], Will Grover[1], and Richard A. Mathies[1]
[1]Department of Chemistry, University of California, Berkeley, CA 94720
[2]Biophysics Graduate Group, University of California, Berkeley, CA 94720
charlie@zinc.cchem.berkeley.edu

Modern microfabrication technologies facilitate the creation of novel genetic analysis devices having dramatically enhanced capabilities. We present here the results of three research projects with respective aims of developing: (i) a 384-channel Capillary Array Electrophoresis (CAE) microdevice for ultra high-throughput genotyping; (ii) an integrated optical detection system using high-sensitivity a-Si:H PIN diodes fabricated on glass; and (iii) a microfluidic-based DNA computer.

Microfabricated devices are replacing conventional drawn silica capillaries for use in high-throughput electrophoresis assays. Toward this end, we have designed and successfully demonstrated a radial 384-lane CAE microdevice and used it to genotype 384 individuals in less than 7 minutes (1). The CAE devices were fabricated on 200-mm glass wafers with channels 50 μm in diameter. Detection was accomplished with our rotary confocal laser-induced fluorescence scanner (2). We demonstrated the efficacy of the device for genotyping by testing 384 individuals for the common H63D mutation in the human *HFE* gene from PCR-RFLP derived samples.

Integration of the fluorescence detector on-chip is a fundamental challenge that will lead to the development of portable point-of-care lab-on-a-chip (LOC) platforms. Conventional semiconductor photodiodes are candidates for integrated detectors, but the required high-temperature fabrication procedures are not compatible with the glass or plastic substrates used for most lab-on-chip devices. We have instead chosen (in collaboration with Xerox PARC) to develop photodiodes made from hydrogenated amorphous silicon (a-Si:H) which can be fabricated at low cost and produced in large

arrays. We have successfully performed DNA fragment sizing using the a-Si:H detector coupled to our current confocal fluorescence station demonstrating the feasibility of miniaturizing and integrating such detectors into a portable LOC device.

Finally, we are developing a microfluidic device to serve as a programmable microprocessor for execution of DNA-based computation. Our current device is an orthogonal array of 81 interconnected chambers containing structures that capture, release, or redirect populations of oligonucleotides that flow through the chambers. Such a device was first postulated and modeled theoretically in 1999 by Gehani and Reif (3). We have expanded on their model by incorporating the traditional Boolean logic gates AND, OR, and NOT specified by the path followed by a particular DNA molecule through the microprocessor. Using micromole quantities of input oligonucleotides it should be possible to encode all possible solutions to a complex combinatorial problem or to aid in solving *NP*-hard problems. The DNA microprocessor can also be used as a haplotyping tool to detect linkages across different polymorphisms using partially digested genomic DNA as logical inputs.

1. Emrich, C. A.; Tian, H.; Medintz, I. M.; Mathies, R. A. *in preparation.*
2. Shi, Y. N.; Simpson, P. C.; Scherer, J. R.; Wexler, D.; Skibola, C.; Smith, M. T.; Mathies, R. A. *Analytical Chemistry* **1999,** *71,* 5354-5361.
3. Gehani, A.; Reif, J. *Biosystems* **1999,** *52,* 197-216.

46. Microarray Electrophoretic DNA Mapping System

Gregory Zeltser, Alfred Goldsmith, Ilya Agurok, and Paul Shnitser
Physical Optics Corporation
GZeltser@aol.com

Physical Optics Corporation (POC) proposes to develop a novel Microarray Electrophoretic DNA Mapping (DNA-MEM) system based on a multistage electrophoresis chip to carry out rapid multiple DNA molecule homogeneous stretching and mapping with a resolution below kilobase. This system will bring needed improvements to current DNA optical mapping and DNA fiber FISH technologies especially in terms of resolution.

The DNA-MEM system consists of three major components: the multistage electrophoresis chip, spectroscopic imaging microscope, and laser diode heating subsystem. DNA molecules are stretched and immobilized into the 3-D mesh of the POC's hydrophilic polymer inside grooves of the chip. The DNA-MEM system will be able to process multiple samples at the same time. Moreover, the system can analyze both preliminarily stained DNA molecules suspended and stretched in porous polymer and DNA molecules that have been FISH stained after stretching and immobilization to the three-dimensional meshwork of the polymer. The DNA-MEM system design, analysis, and component development will be completed in Phase I.

Integrated biochip technology will be developed to the prototype stage in Phase II, and will be commercialized over the following two years as a Phase III engineering prototype.

Functional Analysis and Resources

47. Comparative and Functional Genomics Technologies

Robi Mitra, Vasudeo Badarinarayana, John Aach, Wayne P. Rindone, and George C. Church
Lipper Center for Computational Genetics, Department of Genetics, Harvard Medical School
church@arep.med.harvard.edu

Our project focuses on developing cost-effective technologies for determining and computationally comparing data on gene expression and selectable phenotypes generally applicable to microbial and vertebrate genomes relevant to the overall DOE goals. In particular we have developed very high-resolution genome-based arrays capable of sub-gemic dissection of phenotypes, detection of alternative RNAs, and DNA-protein-binding sites. We have developed a method for in situ amplification, sequencing, and long range (multi-kilobase) RNA-splice-typing and DNA haplotyping.

For more information see:
http://arep.med.harvard.edu

48. On Telomeres, Linkage Disequilibrium, and Human Personality

R. K. Moyzis[1,2], D. L. Grady[1], Y.-C. Ding[1], E. Wang[1], S. Schuck[2], P. Flodman[2], M. A. Spence[2], and J. M. Swanson[2]
[1]Department of Biological Chemistry, [2]Department of Pediatrics and the Child Development Center, College of Medicine, University of California, Irvine, Irvine, CA 92715 USA
rmoyzis@uci.edu

Human telomeres end with a stretch of the conserved simple repeat sequence (TTAGGG)n. To capture single-copy human DNA regions linked to telomeres, large telomere-terminal fragments of human chromosomes were cloned using specialized yeast artificial chromosome (YAC) vectors. By contrast, bacterial artificial chromosome (BAC) libraries are not expected to contain sequences extending to the telomere, owing to the absence of restriction sites in (TTAGGG)n, the effects of length associated with the construction of size-selected DNA recombinant clones, and the genomic instability of these regions. By DNA sequencing of cosmid subclones derived from telomere YACs, connection to the working draft human sequence has now been accomplished (Riethman et al., *Nature* **409**, 948-951, 2001; www.genome.uci.edu). Integration with the working draft sequence was confirmed for 32 telomeres (out of the 46 distinct ends), with framework sequence extending to within 250kb-50kb of the physical end of these chromosomes. Subtelomeric sequence structure appears to vary widely, mainly as a result of large differences in subtelomeric repeat sequence abundance and organization at individual telomeres. Many subtelomeric regions appear to be gene-rich, matching both known and unknown expressed genes.

The great variability in subtelomeric regions between individuals has potential biological significance. It is unclear, therefore, if finishing a "single" sequence in these regions has biological meaning. We suggest that extensive population/species sampling will be needed to characterize this variability. We have begun targeting a number of subtelomeric regions for such "high-depth" DNA resequencing/haplotyping. One of our first targets, the dopamine receptor D4 (DRD4) gene, located at the telomere of 11p, yielded surprising results. Associations have been reported of the 7-repeat (7R) allele of the DRD4 gene with both attention deficit/hyperactivity disorder (ADHD) and the personality trait of novelty seeking. This polymorphism occurs in a 48 bp tandem repeat (VNTR) in the coding region of DRD4, with the most common allele containing four repeats (4R), and rarer variants containing two (2R) to eleven (11R) repeats. By DNA resequencing/haplotyping of over 1000 DRD4 alleles, representing a worldwide population sample, we uncovered that the origin of 2R- through 6R-alleles can be explained by simple one-step recombination/mutation events. In contrast, the 7R-allele is not simply related to the other common alleles, differing by greater than 6 recombinations/mutations. Strong linkage

disequilibrium (LD) was found between the 7R-allele and surrounding DRD4 polymorphisms, suggesting this allele is at least 5-10 fold "younger" than the common 4R-allele. Based on an observed bias towards nonsynonymous amino acid changes, the unusual DNA sequence organization, and the strong LD surrounding the DRD4 7R-allele, we propose that this allele originated as a rare mutational event, that nevertheless increased to high frequency in human populations by positive selection (Ding, et. al., *PNAS*, in press, 2001).

49. Strategies for Construction of Subtracted Libraries Enriched for Full-Length cDNAs and for Preferential Cloning of Rare mRNAs

Brian Berger, Sergey Malchenko, Irina Koroleva, Einat Snir, Tammy Kucaba, Maria de Fatima Bonaldo, and **Marcelo Bento Soares**
The University of Iowa, Departments of Pediatrics, Biochemistry, Physiology and Biophysics
bento-soares@uiowa.edu

Subtracted libraries enriched for full-length cDNAs.
A major challenge of the ongoing NIH Mammalian Gene Collection Program is the identification of sufficient novel full-length cDNAs to enable achieving the yearly full- length sequencing goals of the project. In an effort to assist in the identification of novel full-length cDNAs we have constructed full-length-enriched libraries and we have developed a novel method for generation of subtracted libraries enriched for full-length cDNAs. Conventional subtractive hybridization procedures cannot be applied for full- length-enriched libraries because a truncated clone in the driver population has the potential to subtract its full-length counterpart from the library. Briefly, 100-150 bp single-stranded overhangs are generated at the 5' end of all clones in the library (tracer), for hybridization with a biotinylated driver population comprising representative clones of every sequence contig identified in the starting full-length- enriched library. The subtracted population is purified from the hybrids using streptavidin-coated magnetic beads, repaired and electroporated into bacteria for propagation of a subtracted full-length-enriched library. We have used this method successfully to generate a subtracted full-length-enriched library derived from germinal center B cells.

Preferential cloning of rare mRNAs.
Discovery of rare mRNAs in large-scale EST projects remain difficult and inefficient because of poor representation of such transcripts in cDNA libraries. In an attempt to expedite the identification of rare mRNAs, we developed a novel method for prerential cloning of rare mRNAs. Briefly, mRNA is hybridized with a driver comprising most/all already identified cDNAs and subsequently destroyed with RNAse H. The remainder intact mRNA is linearly amplified and cloned for production of a library enriched for rare mRNAs. We have used this method to construct a mouse cDNA library enriched for rare mRNAs from hippocampus. The efficacy of our method was demonstrated by sequencing and by microarray hybridization analyses.

50. The IMAGE Consortium: Moving Toward a Complete Set of Full-Length Mammalian Genes

P. Folta, N. Ghaus, N. Groves, T. Harsch, A. Johnston, P. Kale, C. Sanders, K. Schreiber, and **C. Prange**
Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory
prange1@llnl.gov

The I.M.A.G.E. Consortium comprises the largest publicly available collection of cDNAs; currently encompassing over 5.5 million clones from six species. These clones are arrayed at Lawrence Livermore National Laboratory, sequenced at various centers, and the resulting ESTs are immediately deposited into Genbank. The clones themselves are made available through a network of distributors worldwide. Rearrayed clone sets representing unique genes of interest are also developed and distributed througb the I.M.A.G.E. pipeline.

Over the last 18 months, IMAGE has been involved in arraying and rearraying clones in support of the Mammalian Gene Collection (MGC) project, an NIH-sponsored effort to generate full-length cDNA resources (http://mgc.nci.nih.gov/). Both EST and full-insert sequences are generated from full-length

enriched cDNA libraries. As of November 2001, clones from 100 enriched libraries have been arrayed, resulting in over 1 million ESTs submitted to dbEST. Sequence analysis predicts more than 30,000 of these clones to be unique and full-length. These clones are rearrayed at LLNL and sent to various sequencing centers for full-length sequencing. All sequences generated from the MGC clones are deposited into Genbank, and all clones and rearrayed clone sets are available royalty-free through the I.M.A.G.E. distributors. At this time over 10,000 full-length high-quality human and mouse sequences have been submitted to Genbank.

Another main focus of the I.M.A.G.E. Consortium has been the development of database query tools to aid in the tracking and analysis of clone-related data. These tools offer web-based query capabilities interconnecting many areas of interest, including clones, libraries, tissues, sequences, rearrays, EST and gene clusters, and quality control information. These tools have contributed to the ease of use of this collection and we are continuing to add additional query capabilities.

Further information about the I.M.A.G.E. Consortium is available by email (info@image.llnl.gov) or through the WWW (http://image.llnl.gov).

This work was partially funded by the NIH and was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract no. W-7405-Eng-48.

51. Functional Genomics Research in AIST-JBIRC

Naoki Goshima, Tohru Natsume, Kousaku Okubo, and **Nobuo Nomura**
Japan Biological Information Research Center (JBIRC), National Institute of Advanced Industrial Science and Technology (AIST), Japan
nnomura@jbirc.aist.go.jp

Functional genomics group of JBIRC implements functional analysis of genes and proteins based on 30,000 human full-length cDNA clones which have been collected by collaborators since 1998. There are four teams in the group. The goals of each team are as follows:

i. **Protein Expression Team.** The complete ORF regions of human full-length cDNAs will be cloned in Gateway entry vectors, which are versatile clones for transferring DNA segments to various expression vectors in high throughput.
ii. **Protein Network Team.** The primary objective is to discover potential interacting partners and to establish members of functional protein machinery complex using mass spectrometry. Post-translational modifications regulating protein-protein interactions will be also studied.
iii. **Expression Profiling Team.** Expression profiles and their changes of human genes in cells under both normal and disordered conditions will be quantitatively recorded by the iAFLP method using human full-length cDNA sequence information.
iv. **Cellular Function Team.** Gene function will be studied by introduction of expression cDNA clones into cells.

The high throughput system, which will quantitatively detect the morphological change of cells including processing, bowing, enlargement and others, will be developed.

52. The *Drosophila* Gene Collection

Mark Stapleton[1], Peter Brokstein[2], Guochun Liao[2], Ling Hong[2], Mark Champe[1], Brent Kronmiller[1], Joanne Pacleb[1], Ken Wan[1], Charles Yu[1], Joe Carlson[1], Reed George[1], Susan Celniker[1], and Gerald M. Rubin[2]
[1]Lawrence Berkeley National Laboratory, BDGP, Berkeley, CA
[2]Howard Hughes Medical Institute, University of California Berkeley, Berkeley, CA
staple@bdgp.lbl.gov

The Berkeley *Drosophila* Genome Project's future goals are in functional genomics. Taking advantage of the *Drosophila* genome sequence, we intend to develop tools and technologies for answering biological questions in a high-throughput environment. Our first step in this direction is to create a publicly available collection of *Drosophila* cDNAs, sequence them to high quality, and begin

converting them into universal Gateway (LifeTechologies) clones. Using an in vitro recombination reaction based on phage lambda, Gateway clones can be subcloned en masse into a variety of expression vectors. In a pilot experiment using Gateway technology, we have created 72 Baculovirus expression constructs representing 36 *Drosophila* transcription factors. Release 1.0 of the *Drosophila* Gene Collection (DGC) has been described (Rubin et al. *Science* 2000). It was produced by sequencing some 80,000 5' ESTs from cDNA libraries derived from various tissues and stages. The DGC Release 1 consists of a non-redundant set of nearly 6,000 clones – 42% of all predicted genes. We are currently full-insert sequencing Release 1 using a commercially available in vitro transposition system. Data will be presented for full-insert sequencing utilizing this transposon-based methodology. Since the DGC Release 1 comprises a fraction of the predicted genes in *Drosophila*, we have generated an additional 160,000 5' ESTs from existing and newly constructed libraries. The new libraries were generated in a collaboration with Piero Carnicci at the RIKEN in Japan. Given the availability of a highly annotated genome sequence, we have computationally selected over 5,000 clones to generate DGC Release 2, which contains over 11,000 clones. BDGP now has clones representing almost 75% of all predicted genes in *D. melanogaster*. Release 2 has been added to our full-insert sequencing pipeline and should be completed in early 2002. We are now focusing on identifying and sequencing major splice forms as well as developing directed approaches to obtain full-length cDNAs for the remaining genes.

53. Identification of the Complete Regulon of a Master Transcriptional Regulator

Michael Laub, **Swaine Chen**, Lucy Shapiro, and Harley McAdams
Department of Developmental Biology, Stanford University
slchen@stanford.edu

The objective of the Stanford Microbial Cell project is to identify the complete transcriptional regulatory network of the aquatic bacterium, *Caulobacter crescentus*. In this poster, we describe how we have applied a combination of experimental and bioinformatic techniques to determine the complete regulon controlled by CtrA, a master transcriptional regulator that controls many *Caulobacter* cell cycle processes including DNA replication, polar morphogenesis, and cell division. We used an in vivo technique involving cross-linking bound CtrA to its binding site followed by fragmenting the DNA and using immunoprecipitation to enrich the segments with linked CtrA proteins. We then reversed the crosslinks and used a microarray assay to identify the enriched DNA segments. We combine this binding site assay with data on RNA expression patterns in wild type and mutant cells to determine the complete CtrA regulon.

54. Deciphering the Gene Regulatory Network of a Simple Chordate

Byung-in Lee[1], David Keys[2], Andrae R. Arellano[1], Chris J. Detter[1], Paul Richardson[1], Michael Levine[1,2] Mei Wang[1], Orsalem J. Kahsai[1], David K. Engle[1], Irma Rapier[1], Sylvia Ahn[1] and Trevor Hawkins[1]
[1]DOE Joint Genome Institute, Walnut Creek, CA 94598
[2]Department of Molecular and Cellular Biology, University of California at Berkeley, Berkeley, CA 94720
lee110@llnl.gov

Regulatory DNA elements such as promoters and enhancers work by serving as docking sites for specific protein complexes. These complexes are comprised of cooperative groups of transcription factor proteins that recognize the target DNA sequences quite specifically and their presence or absence governs the off or on status of their target regulatory sites. Therefore an understanding DNA regulatory element is to understand the composition and function of the biochemical networks and pathways that carry out the essential processes of living organisms.

To characterize gene regulatory network, we used electroporation assays to screen genomic DNA fragments for tissue specific enhancer activities in *Ciona intestinalis*. The *Ciona* genome is one of the smallest and most compact of all chordate genome and *Ciona* tadpole represents the most simplified chordate body plan (the ciona notochord contain only 40 cells). Since the synchronously developing

embryos from *Ciona* can be introduced a transgenic DNA via simple electroporation, we determine *cis* regulatory DNA modules that lead to the specification of each of the key chordate tissues.

We screened ~300kb of BAC DNA which contained HOX gene clusters for tissue specific enhancer elements using the shotgun approach, and found 37 screened clones (80kb) of positive tissue specific enhancer elements. And 13 different tissue types were recognized from the screening. Currently we are investigating minimum enhancer elements from random genomic pieces and also whole mount *in-situ* hybridization with genes within.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, the Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098, and the Los Alamos National Laboratory under contract No. W-7405-ENG-36.

55. Functional Analysis of Gene Regulatory Networks Underlying Skin Biology and Environmental Susceptibility

Brynn H. Jones[1], Jay R. Snoddy[1], Cymbeline T. Culiat[1,3], Mitchel J. Doktycz[1,3], Peter R. Hoyt[1], Denise D. Schmoyer[2], Erich J. Baker[1], Douglas P. Hyatt[1], Line C. Pouchard[2], Michael R. Leuze[2], Eugene M. Rinchik[1,4], and Edward J. Michaud[1,3]
[1]Life Sciences Division, and [2]Computer Science and Mathematics Division, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831
[3]The University of Tennessee–Oak Ridge National Laboratory Graduate School of Genome Science and Technology, Oak Ridge, TN 37830
[4]Department of Biochemistry, Cellular, and Molecular Biology, The University of Tennessee, Knoxville, TN 37996
jonesbh@ornl.gov

Deciphering the complex biological systems that underlie human health and susceptibility to the environmental consequences of energy production is an important mission of the DOE's Biological and Environmental Research Program. For many years, ORNL has focused on annotating human DNA sequence information with gene function information based on the genetic analysis of induced single gene mutations in the mouse (see abstracts by Michaud et al., and Culiat et al.). This single-gene, functional-genomics approach leads naturally to a parallel dissection of complex gene regulatory networks, and of the role of individual genetic variation in susceptibility to environmental agents. The recent availability of the complete genomic sequences from humans and mice, and new technologies to assess gene regulation in a high-throughput manner has dramatically increased our ability to elucidate complex biological systems. Here we describe a new project that combines three areas of expertise at ORNL (mouse molecular genetics, analytical technologies and instrumentation, and bioinformatics and computational biology), designed to develop an integrated-systems approach for defining gene function in skin biology and environmental susceptibility. Our initial efforts focus on a novel Oak Ridge mutation (Hrn) in a transcription factor encoded by the hairless (hr) gene. Hairless mutants are characterized by early and persistent loss of body hair, and by increased susceptibility to UV- and chemical-induced carcinogenesis, and to dioxin toxicity. Using skin-specific cDNA microarrays we have identified numerous genes that are differentially expressed in the skin of Hrn mutants, thus identifying some of the components of the regulatory network associated with the hairless transcription factor. In parallel we are applying the concept of phylogenetic footprinting to the task of elucidating this gene regulatory network. Data obtained experimentally with hairless mutants will be used as the empirical basis for understanding co-regulation of gene expression using bioinformatics tools. Ultimately we will be able to predict membership of genes in networks based on the presence of shared binding site motifs in regulatory regions, and these hypothesis may be tested by examining the whole-animal consequences of induced mutations in the regulatory sequences and coding regions of each network component using ORNL's Cryopreserved Mutant Mouse Bank (CMMB).

[Research sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, and by the Office of Biological and Environmental Research, U.S. Department of Energy under contract DE-AC05-00OR22725 with UT-Battelle, LLC.]

56. Genomic Identification and Analysis of Shared cis-Regulatory Elements in a Developmentally Critical Homeobox Cluster

Tsutomu Miyake, Mark Dickson, Jane Grimwood, Steve Irvine, Andrew Brady Stuart, Jeremy Schmutz, Kenta Sumiyama, Richard M. Myers, Frank H. Ruddle, and **Chris T. Amemiya**
Virginia Mason Research Center, Stanford University School of Medicine, Yale University
camemiya@vmresearch.org

A major problem in biology is the delineation of how a one-dimensional sequence of nucleotides can specify a three-dimensional organism. Central to this process is the assurance that the information hardwired into the DNA sequence directs the regulation of genes in their proper temporal and spatial milieu. This coordinated regulation of genes is very complex, and underlies all biological processes, including development, differentiation, evolution, speciation, and the onset of disease. Numerous studies have been performed using the relatively laborious method of site-directed mutagenesis and subsequent expression analysis, in order to deduce the identity and nature of *cis*-regulatory elements (enhancers and repressors) at a fundamental level. However, alternative experimental approaches are clearly required to detect and characterize these sequences in order to better understand the systematic and interactive roles they play, particularly in a more global context. This is especially true of developmentally important "gene complexes" which are regulated in a programmatic fashion during development. The fact that the structure, organization, and developmental expression patterns of these genes have been so strikingly conserved throughout metazoan evolution, suggest that there exists sequence-encoded mechanisms ensuring their evolution and deployment *in concert*. We propose using a combination of genomic, molecular, cellular and morphologic tools in order to make inroads into our understanding of this problem. We will focus our attention on the *Distalless* (*Dlx*) homeobox clusters, whose developmental significance is well established with respect to pattern formation. These relatively small gene clusters serve as regulatory models for other developmentally critical gene clusters in complex vertebrate genomes (such as the *Hox* clusters, olfactory receptors, and genes of the anticipatory immune system). We will incorporate a highly integrated approach for the comparative analysis of these clusters among selected mammalian taxa. This pilot project will necessarily implement aspects of genomics (BAC analysis, long-range DNA sequencing, bioinformatics/ computation) and developmental biology (transgenic and knockout/knock-in technologies). The broad goals of this project are to identify and understand the genomic basis for the cooperative regulation of the *Distalless* (*Dlx7/Dlx3*) genes, and to further develop this experimental paradigm for future, larger-scale studies of genomic control. In addition to the practical biological/medical implications of this pilot study, the dataset should prove highly useful for evaluation of novel computational methods for multiple sequence alignments.

57. A Sequence-Ready Comparative Map of Chicken Genomic Segments Syntenically Homologous to Human Chromosome 19

Laurie Gordon, Joomyeong Kim, Hummy Badri, Mari Christensen, Matthew Groza, Mary Tran, and Lisa Stubbs
DOE Joint Genome Institute, 2800 Mitchell Drive, Building 100, Walnut Creek, CA 94598-1604 and Genomics Division, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, 7000 East Avenue, L-441, Livermore CA 94550
gordon2@llnl.gov

Having recently completed mapping and sequencing mouse genomic segments syntenically homologous to human chromosome 19 (HSA19), we are generating a parallel set of sequence-ready chicken BAC clone contigs. The locations of some HSA19-homologous genes are known in chicken, but homology segments are not well characterized. Preliminary comparisons of human, mouse and chicken conserved segments by other groups suggest that the organization of the chicken genome is more like that of human than mouse. Comparative sequencing of a third vertebrate with greater evolutionary distance from the two mammalian species will test and expand these findings,

facilitating a better understanding of ancestral chromosomal organization and gene evolution.

Protein-translated HSA19 gene sequences identified well conserved (60-95%) chicken ESTs for ≈160 gene loci. Overgo and PCR probes were developed wherever conserved sequences were identified to facilitate detection of gene synteny and homology segment breakpoints. Probes were hybridized to two BAC libraries, one each from *Gallus domesticus* and *Gallus gallus* (5x, respectively). Clones identified by hybridization were restriction digested and assembled into maps, generating important information on clonal integrity, length and overlap; restriction maps also facilitate contig extension, identification of potential joins between neighboring contigs and sequencing tiling path selection. To date we have successfully identified at least one bac for ≈80 gene loci, generating forty-five contigs covering 7 Mb of the chicken genome. Given the extraordinary compactness of the chicken genome relative to human, we estimate current coverage of approximately 40% of euchromatic HSA19-related territory. We have confirmed the presence of syntenic homology segments while detecting significant rearrangements relative to human and mouse, including differential organization of clustered gene families. Representative bacs are being sequenced by the Joint Genome Institute in Walnut Creek; comparative data will facilitate the characterization of HSA19-related homology segments and shed light on chromosomal evolution in vertebrates.

58. Characterization of a New Imprinted Domain Located in Human Chromosome 19q13.4/ Proximal Mouse Chromosome 7

Joomyeong Kim, Anne Bergmann, Edward Wehri, Xiaochen Lu, and Lisa Stubbs
Genomics Division, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94550
Kim16@llnl.gov

For a subset of mammalian autosomal genes, the two parental alleles are not functionally equivalent due to genomic imprinting. Imprinting involves inactivation of one allele, depending upon the parental origin. More than 40 imprinted genes have been identified from human and mouse in the past 10 years. Most imprinted genes are thought to be involved in either fetal growth or animal behavior, and most imprinted genes are found clustered in specific regions of chromosome, suggesting the presence of long-range regulatory mechanisms for genomic imprinting. In early studies, we located one imprinted gene, Peg3 (paternally expressed gene 3), to human chromosome 19q13.4. Due to the clustering of imprinted genes in specific chromosomal regions, it seemed likely that other imprinted genes would be found in the interval surrounding PEG3. We have since isolated and characterized most genes located in the 1MB-genomic intervals surrounding human and mouse PEG3. Our studies have identified six new imprinted genes in this new domain, including Peg3, Zim1 (imprinted Zinc-finger gene 1), Zim2, Zim3, Usp29 (Ubiquitin-specific processing protease 29), and Znf264. Most of these new imprinted genes are predicted to function as transcription factors based on the zinc-finger motifs detected in the predicted proteins of these genes. In contrast to most imprinted regions, the HSA19q and Mmu7 imprinted domains have changed considerably in terms of the content, coding capacity and transcriptional activities of resident genes.

Two different directions are currently being developed in our lab for the future study. First, we are working to characterize the physiological functions of these new imprinted genes using mouse genetic approaches. Second, we are using comparative genomics approaches to study the regulatory mechanism controlling the imprinting and expression of these six genes. Based on our preliminary results, it is likely that one region, the surrounding region of the first exon of Peg3, might be responsible for the imprinting of a whole domain and we are presently aiming to test this hypothesis as well as to identify regulatory regions associated with all 6 imprinted genes.

59. A New Apolipoprotein Influencing Plasma Triglyceride Levels in Humans and Mice Revealed by Comparative Sequence Analysis

Len A. Pennacchio[1], Michael Olivier[3], Jaroslav A. Hubacek[2], Jonathan C. Cohen[2], Ronald M. Krauss[1], and Edward M. Rubin[1]
[1]Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 USA
[2]Center for Human Nutrition and McDermott, Center for Human Growth and Development, UT Southwestern Medical Center, Dallas, TX 75390-9052 USA
[3]Human and Molecular Genetics Center, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226 USA
LAPennacchio@lbl.gov

The apolipoprotein gene cluster on human chromosome 11q23 (ApoAI/CIII/AIV) is a well-studied genomic interval that influences a variety of plasma lipid parameters and atherosclerosis susceptibility in humans. To facilitate the identification of evolutionarily conserved sequences with potential function near this cluster, we determined the sequence of ~200 kilobasepairs (kbp) of orthologous mouse DNA and compared the mouse and human sequences. The presence of a stretch of inter-species sequence conservation approximately 30 kbp proximal to the ApoAI/CIII/AIV gene cluster, led us to an interval that upon further analysis was shown to encode a new member (ApoAV) of the chromosome 11 apolipoprotein gene cluster. We find that the ApoAV gene is expressed primarily in liver tissue and encodes a secreted protein that dramatically impacts plasma triglyceride levels in humans and mice. Specifically, mice over-expressing a human ApoAV transgene display a 70% decrease in plasma triglyceride concentrations, while oppositely, mice lacking ApoAV have a 400% increase in this lipid parameter. These findings in mice suggested that alterations in ApoAV could also influence human plasma lipid levels. To explore this possibility, we identified several single nucleotide polymorphisms (SNPs) in the human ApoAV gene and determined their distribution in two independent patient populations. Through this analysis, we found a significant association between several polymorphisms and abnormal triglyceride levels in both independent studies. Heterozygous individuals had on average a 32% increase in plasma triglyceride levels when compared to individuals homozygous for the common allele. We determined that approximately 20% of the Caucasian population contain an ancestral chromosomal fragment representing a definable susceptibility haplotype. These findings in humans and mice illustrate the utility of comparative sequence analysis to prioritize regions of the genome for further study and suggest an important physiological role for ApoAV in affecting plasma levels of triglyceride, a major risk factor for heart disease in humans.

60. *Nell1*: A Candidate Gene for ENU-Induced Recessive Lethal Mutations at the *l7R6* Locus and Potential Mouse Models for Human Neonatal Unilateral Coronal Synostosis (UCS)

Cymbeline T. Culiat[1], Jennifer Millsaps[2], Jaya Desai[3], Beverly Stanford[1], Lori Hughes[1], Marilyn Kerley[1], Don Carpenter[1], and Eugene M. Rinchik[1, 3]
[1]Life Sciences Division, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, TN 37831-8077
[2]Genome Science and Technology Graduate School, The University of Tennessee, Knoxville, TN 37996
[3]Department of Biochemistry, Cellular and Molecular Biology, The University of Tennessee, Knoxville, TN 37996

A gene (*l7R6*) critical for late embryonic development and survival has been mapped proximal to the pink-eyed dilution (*p*) gene in mouse chromosome 7. Six independent ENU-induced alleles designated 88SJ, 335SJ, 2038SJ, 102DSJ, 11DSJ and 45DSJ all result in late-gestation/ neonatal lethality. *l7R6* maps to a region homologous to human 11p15.1, that contains a very large gene for a protein kinase C binding protein, called *NELL1*. Human *NELL1* has a 2433-bp coding region with at least 20 exons spread out in ~800 kb genomic distance. Because the human gene is so large, and because we recovered so many *l7R6* alleles in a relatively small number of gametes, the mouse counterpart seemed a logical candidate for *l7R6*. To determine if *l7R6* is *Nell1*, a near full-

length (1920 bp) cDNA was used as a probe for Northern analysis of both wild-type and mutant animals. *Nell1* expression was detected from E10-E18, increasing as fetal development progresses and concentrating particularly in the head at E18. In wild-type adults, expression was predominantly in brain. Notably, a severely reduced level of *Nell1* expression was detected in one allele (102DSJ). Abnormal expression of human *NELL1* is associated with unilateral coronal synostosis (UCS) in newborns, a condition where coronal sutures fuse early, resulting in abnormal head development and limb defects. Mouse hemizygotes recovered at either E18 or two hours after birth also exhibit both gross cranial and limb defects. Cloning and sequencing of RTPCR-derived cDNA clones from mutant and wild-type alleles have shown that the mutation in the 335SJ allele is an AT/GC transition resulting in a cysteine to arginine substitution in the *Nell1* protein. The phenotype data, RNA analysis and mutation scanning experiments indicate that *l7R6* is *Nell1*. Studying the spectrum of mutations in this allelic series will be valuable in understanding the structure and function of the *Nell1* protein.

[Research sponsored by the Office of Biological and Environmental Research, U.S. Department of Energy under contract DE-AC05-00OR22725 with UT-Battelle, LLC.]

61. Functional Annotation of Human Genes with Phenotype-Driven and Gene-Driven Mutagenesis Strategies in Mice

Edward J. Michaud[1,2], Carmen M. Foster[1], Rosalynn J. Miltenberger[1,2], Miriam L. Land[1], Dabney K. Johnson[1,2], and Eugene M. Rinchik[1,3]
[1]Life Sciences Division, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831-6445
[2]The University of Tennessee-Oak Ridge National Laboratory Graduate School of Genome Science and Technology, Oak Ridge, TN 37830-8026
[3]Department of Biochemistry, Cellular, and Molecular Biology, The University of Tennessee, Knoxville, TN 37996
michaudejiii@oml.gov

One focus of the Mouse Genetics and Genomics Program at ORNL is to determine the whole-organism biological functions of human genes by inducing mutations in the homologous genes in mice. Our program currently uses a phenotype-driven chromosome-region mutagenesis strategy in the mouse to identify and map gene function in pre-selected segments of the genome, which together total approximately 8% of the mouse genome. The genetic reagents that are necessary to perform these chromosome-region mutagenesis screens, however, are not currently available for most of the mouse chromosome regions that are homologous to the human chromosomes (5, 16, and 19) sequenced by the DOE. In this project, we are exploiting newly developed techniques for engineering chromosomes in mouse embryonic stem cells that will permit phenotype-driven mutagenesis screens and functional-genomics analyses to be performed in any region of the genome. Specifically, we are generating radiation-induced deletions and Cre-loxP-mediated inversions in large, gene-rich regions of mouse chromosomes that are in synteny conservation with human chromosomes. Our initial focus is the proximal two-thirds of mouse chromosome 7, which has homology to all of human chromosome 19q and to portions of chromosomes 11p, 15q, and 11q. Although chromosome-region phenotype-driven mutagenesis in mice is currently the state-of-the-art for identifying and mapping the biological functions of human genes, engineering the appropriate genetic reagents for a new chromosome region and performing the mutagenesis screens are still time consuming endeavors. To augment our phenotype-driven mutagenesis strategy, we are taking advantage of the recent availability of the draft sequence of the mouse genome to develop a new DNA sequence-driven or gene-driven mutagenesis strategy (ORNL's Cryopreserved Mutant Mouse Bank, CMMB; see other abstract by Michaud et al.) that will allow us to determine the biological functions of any pre-selected genes in the genome, regardless of their chromosomal locations. The ongoing development of the CMMB offers us a new and unprecedented opportunity to apply our expertise in chemical germ-cell mutagenesis in mice specifically to understanding the biological functions of those genes on human chromosomes 5, 16, and 19.

[Research sponsored by the Joint Genome Institute, Office of Biological and Environmental Research, U.S. Department of Energy under contract DE-AC05-00OR22725 with UT-Battelle, LLC.]

62. Resource Archiving and Distribution via the Mutant Mouse Database and the Cryopreservation Program at the Oak Ridge National Laboratory

D. K. Johnson[1], E. M. Rinchik[1,2], P. R. Hunsicker[1], S. G. Shinpock[1], K. J. Houser[1], D. J. Carpenter[1], G. D. Shaw[1], W. Pachan[1], **E. J. Michaud**[1], B. L. Alspaugh[3], and L. B. Russell[1]
[1]Life Sciences Division, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, TN 37831-8077
[2]Department of Biochemistry, Cellular and Molecular Biology, The University of Tennessee, Knoxville, TN 37996
[3]Science Applications International Corporation, Oak Ridge, TN 37830
v71@ornl.gov

The Mouse Genetics and Genomics Program at Oak Ridge National Laboratory (ORNL) currently curates eight hundred standard or mutant strains of laboratory mice in a conventional colony co-located with laboratory space equipped for molecular biology, broad-based phenotype screening, and genetic engineering. Of these 800 strains, 300 are in live maintenance and 670 are banked as cryopreserved embryos, sperm, and/or ovaries. Detailed information about actively propagated or cryopreserved stocks is listed in ORNL's searchable Mutant Mouse Database (http://bio.lsd.ornl.gov/mouse/). Mutant stocks may be obtained as breeding pairs, frozen tissues, or frozen embryos, sperm, or ovaries for a cost-recovery fee. Mouse stocks are cryopreserved as embryos or germ cells in order to archive stocks not in current active use, to preserve the necessary materials for the rederivation of all stocks into our new barrier facility, and to provide a means for distribution of requested stocks to the international research community. Protocols and progress of the cryopreservation effort may be viewed at the Mammalian Genetics and Genomics Program website (http://bio.lsd.ornl.gov/mgd), which includes a further link (http://tnmouse.org/) to information on ORNL's participation in the Tennessee Mouse Genome Consortium, one of the national mouse-mutagenesis centers funded by the National Institutes of Health. ORNL has recently begun construction of the William L. and Liane B. Russell Laboratory of Comparative and Functional Genomics, a 30,000 square-foot colony built to house 60,000 mice in specific pathogen-free conditions. This new, SPF mouse-breeding facility will be important for the future research activities of the ORNL's Mouse Genetics and Genomics Program.

[Research sponsored by the Office of Biological and Environmental Research, U.S. Department of Energy under contract DE-AC05-00OR22725 with UT-Battelle, LLC.]

63. Mutation Scanning and Candidate-Gene Verification in the ORNL Regional ENU-Mutagenesis Program

Cymbeline Culiat[1], Qingbo Li[2], Mitchell Klebig[3], Dabney Johnson[1], Zhaowei Liu[2], Heidi Monroe[2], Beverly Stanford[1], Tse-Yuan Lu[1], Lori Hughes[1], Marilyn Kerley[1], Don Carpenter[1], Lisa Webb[4], and Eugene M. Rinchik[1,3]
[1]Life Sciences Division, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, TN 37831-8077
[2]SpectruMedix Corporation, 2124 Old Gatesburg Rd, State College, PA 16803
[3]Department of Biochemistry, Cellular and Molecular Biology, The University of Tennessee, Knoxville, TN 37996
[4]Graduate School for Genome Sciences and Technology, , The University of Tennessee, Knoxville, TN 37996
9c9@ornl.gov

Identifying the gene alterations in mouse mutations and understanding the resulting perturbed pathways contribute to the functional annotation of the corresponding genes in the human genome. Regional mutagenesis efforts at Oak Ridge National Laboratory have generated and fine-mapped 15 recessive-lethal N-ethyl-N-nitrosourea (ENU)-induced mutations to a small genomic region proximal to the pink-eyed dilution (p) gene in mouse chromosome 7. These mutations represent six genes important for early mammalian development and

survival. Candidate genes were assigned to these mutations based on integrating data from genetic and physical mapping, phenotype characterization (e.g., time of death studies), expression profiling (regional transcriptomics), and utilization of publicly available bioinformatics data. Three mutation-scanning techniques (dHPLC/TMHA, TGCE and DNA sequencing) were utilized to identify potential mutations in the candidate genes assigned to the recessive lethal mutations. The application of temperature gradient capillary electrophoresis (TGCE), a new high-throughput heteroduplex analysis technique, permitted rapid assignment of positions where ENU-induced base pair changes were located and is the first demonstration of the effectiveness of using TGCE to find ENU-induced mutations (i.e., SNPs) in the mouse genome. Mutation-scanning data for identification of mutations in the Ldh1 (1 allele), Saa3 (2 alleles), Prmt3 (2 alleles), and Nell1 (8 alleles) genes will be presented. Our data demonstrate that TGCE is an excellent approach for examining several candidate genes for a single mutation or a single gene for a cluster of mutations. In addition, optimization of TGCE protocols for mutation scanning and its success in the regional mutagenesis program have led to its further application in a gene-driven approach for finding mouse mutations genome-wide in the CMMB (Cryopreserved Mouse Mutant Bank) (see abstract by Michaud et al).

[Research sponsored by the Office of Biological and Environmental Research, U.S. Department of Energy under contract DE-AC05-00OR22725 with UT-Battelle, LLC.]

64. Genome-Wide, Gene-Driven Chemical Mutagenesis for Functional Genomics: The ORNL Cryopreserved Mutant Mouse Bank

E. J. Michaud[1,2], J. R. Snoddy[1], E. J. Baker[1], Y. Aydin-Son[2], D. J. Carpenter[1], L. L. Easter[1], C. M. Foster[1], A. W. Gardner[1], K. S. Hamby[1], K. J. Houser[1], K. T. Kain[1], T.-Y. S. Lu[1], R. E. Olszewski[1], I. Pinn[1], G. D. Shaw[1], S. G. Shinpock[1], A. M. Wymore[1], D. K. Johnson[1,2], C. T. Culiat[1,2], E. M. Rinchik[1,3]

[1]Life Sciences Division, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831

[2]The University of Tennessee–Oak Ridge National Laboratory Graduate School of Genome Science and Technology, Oak Ridge, TN 37830

[3]Department of Biochemistry, Cellular, and Molecular Biology, The University of Tennessee, Knoxville, TN 37996

michaudejiii@ornl.gov

A major challenge following the sequencing of the human genome is to determine the biological functions of the estimated 40,000-70,000 genes, and the manner in which these genes are coordinately regulated and affected by environmental factors. By inducing mutations in mouse genes and determining the consequences of the mutations in the whole animal, we gain insight into the functions, regulatory networks, and gene-environment interactions of the homologous human genes. The recent availability of the complete DNA sequence of the mouse genome and high-throughput methods for rapid detection of single-nucleotide polymorphisms (SNPs) has facilitated genome-wide, gene-driven approaches to germline mutagenesis. Gene-driven mutagenesis strategies allow one to perform whole-genome mutagenesis, and then screen for alterations in any pre-selected gene(s) in the genome. To augment embryonic stem-cell-based gene-driven mutagenesis resources, such as gene-trap libraries and banks of N-ethyl-N-nitrosourea (ENU)-mutagenized cells, we are generating a bank of DNA, tissues (for RNAs and proteins), and sperm from 5000 individual C57BL/6JRn mice that carry a load of paternally induced ENU mutations. This ORNL Cryopreserved Mutant Mouse Bank (CMMB) will be a source of induced, heritable SNPs in the regulatory regions

and coding sequences of virtually every gene in the genome. High-throughput Temperature Gradient Capillary Electrophoresis (see abstract by Culiat et al.) is being used to identify mutations in pre-selected genes in the DNAs and RNAs from the CMMB, and mutant mice will be recovered from frozen sperm to determine the biological functions of the homologous human genes. Thus, ORNL's CMMB will provide mouse models of a wide range of altered proteins for phenotypic, gene/protein-network, and structural biology-type analyses. We envision the CMMB as a core component in the integration of mouse mutagenesis, gene-expression microarrays, proteomics, and computational biosciences at ORNL for the purpose of (1) determining the function of every gene on the three human chromosomes sequenced by the DOE (see other abstract by Michaud et al.), (2) deciphering complex biological systems underlying human health and susceptibility to the environmental consequences of energy production (see abstract by Jones et al.), and (3) strategically positioning ORNL to respond to DOE's Genomes to Life program.

[Research sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, and by the Office of Biological and Environmental Research, managed by UT-Battelle, LLC for the U. S. Department of Energy under Contract No. DE-AC05-00OR22725.]

65. Filtering Out Functional Open Reading Frame Fragments from DNA

P. Zacchi[1], D. Sblattero[2], R. Marzari[2], and **A. Bradbury**[3]
[1]CIB, Area di Ricerca, Padriciano 99, Trieste, Italy
[2]Dipartimento di Biologia, Universita' di Trieste, Trieste, Italy
[3]Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545
amb@lanl.gov

In any functional analysis of the protein products of a genome, some method is required to physically isolate open reading frames, as opposed to merely identify them. There are two general approaches to this. In the first method, open reading frames are identified by the application of informatic methods to cDNA, EST, genomic sequences and specific primers are designed to amplify the open reading frame from a suitable source (e.g. cDNA). This can then be cloned into a vector of interest. In the second method, suitable DNA is fragmented and sampled at random, and open reading frames are selected. The first method provides full length open reading frames, while the second provides open reading frames which are fragments of full genes. The first method is relatively time consuming, but is more useful for the study of protein function, while the second method can be carried out more easily, and is more applicable to the study of immunological epitopes within gene products. In a model system, we have applied an example of the second method to the analysis of a monoclonal antibody epitope found in tissue transglutaminase (tTG). The gene for tTG was cloned into an expression plasmid and the plasmid fragmented into fragments of 300bp. This represents a model system in which four genes (tTG, rop, lacI and kanamycin) are present with an approximately equal amount of non coding sequence. The fragments were cloned into a vector designed to select open reading frames and a number of clones were identified which expressed the known mAb epitope. Furthermore, sequencing of random fragments revealed that the selection vector had a strong bias for real open reading frames of known function, and selected few open reading frames of no known biological function. This system is likely to be applicable to the efficient selection of random open reading frames representing the immunological coding potential of single genes, whole micro-organisms, normalized cDNA libraries or collections of cloned open reading frames.

Presentation of this poster is subject to completion and submission of the corresponding patent.

66. Towards High Throughput Antibody Selection

Jianlong Lou[1], Roberto Marzari[2], Peter Pavlik[3], Milan Ovecka[3], Nileena Velappan[3], Leslie Chasteen[3], Vittorio Verzillo[3], Federica Ferrero[6], Daniel Pak[4], Morgan Sheng[4], Chonglin Yang[5], Daniele Sblattero[2], and **Andrew Bradbury**[3]
[1]Dept of Anesthesia 3s50, San Francisco General Hospital, UCSF, San Francisco, California
[2]Department of Biology, Universita' di Trieste, Trieste, Italy

[3]Biosciences Division, Mail Stop M888, Los Alamos National Laboratory, Los Alamos, New Mexico
[4]Department of Neurobiology, Massachusetts General Hospital, Boston, Massachusetts
[5]Department of Biochemistry and Molecular Biology, University of Maryland at Baltimore, Baltimore, Maryland
[6]SISSA, Trieste, Italy
amb@lanl.gov

Phage antibody libraries represent a relatively easy way to generate antibodies against a vast number of different ligands. Although in principle, phage antibody selection should be amenable to automation, this has not yet been described and present selection protocols are far from high throughput. We have reduced phage antibody selection to a microtitre format, and compared selection using this format to traditional selection.

Antibodies were selected against eleven different antigens using either a microtitre plate selection method (using pins rather than wells) or the "traditional" immunotube method. We find that the two methods tend to select different antibodies, with only 10% of antibodies in common, even if the plastic, the antigen and the library used are identical. This is in contrast to the use of the same method to select antibodies, when over 30% of antibodies selected are in common.

We are presently working on automating the phage antibody selection and screening method using a Tecan Genesis workstation and a Qbot picking robot. Results will be presented.

67. A Pilot Project for Identifying and Characterizing Protein Complexes

Edward C. Uberbacher, Frank Larimer, Bob Hettich, Greg Hurst, Michelle Buchanan, Dong Xu, and Ying Xu
Oak Ridge National Laboratory
ube@ornl.gov

This pilot project is developing the central technologies needed to build a Protein Complex Factory (PCF) designed to meet the needs described in Goal 1 of the Genomes to Life program for large scale identification and characterization of protein machines. This facility will eventually be able to identify the protein components of protein machines from microbial and eukaryotic genomes at high throughput and from whole cells. The identification of complexed proteins and the interrogation of protein complex organization will be conducted using a combination of mass spectrometry, protein crosslinking, and computing. The goals of this pilot project are to (i) to develop a combined experimental and computational capability for the identification and interrogation of protein-protein interactions and (ii) to apply the developed methodology to several isolated protein complexes and protein complexes within cell extracts, to demonstrate that the methodology is sufficiently accurate, informative and scalable to whole microbial and eukaryotic cells. The approach utilizes specialized affinity tagged crosslinking reagents capable of crosslinking proteins in complexes and which then allow for separation of crosslinked proteins in a cell extract from non-crosslinked products. A subsequent liquid chromatography and tandem mass spectrometry step can then resolve this mix into distinguishable fragments, and rapidly and selectively interrogate each fragment to obtain a unique sequence mass fragment fingerprint. This fingerprint can be used to identify the proteins involved in interprotein crosslinking through a database search methodology that also provides the identity of the specific amino acids involved in each crosslink. Once identified, the crosslinked positions and other information, such as crosslinker length, can be used as constraints to derive information about the geometry of each protein complex.

Several demonstrations are being developed as proof of principle based on well characterized complexes: (1) mouse wild-type hemoglobin and mutant forms, (2) the GroELS complex, and (3) the pmf ATPase molecular engine.

For these targeted complexes, the pilot project will (a) generate the necessary crosslinked complexes (b) detect and interrogate significant numbers of intermolecular crosslinked fragments by tandem mass spectrometry, and (c) deconvolute and interpret the data in terms of complex identification and organization. As part of the pilot project, estimates will be obtained that directly address issues of scale,

including what data collection will cost and how long it would take to comprehensively examine protein machines in a whole cell. If successful, this pilot will set the stage for a Phase II production capability.

68. High-Throughput Protein Expression and Purification for Proteomics Research

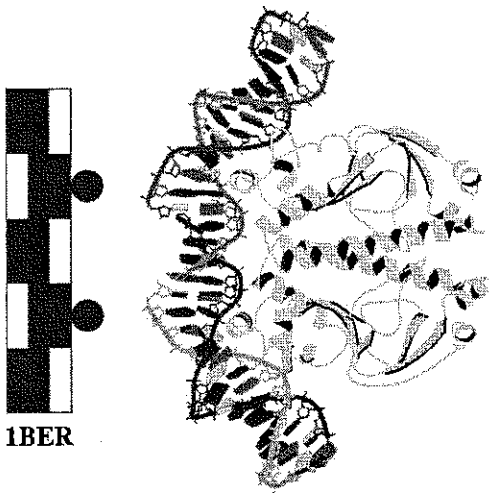
Sharon Doyle, Jennifer Primus, Michael Murphy, Paul Richardson, and Trevor Hawkins
DOE Joint Genome Institute, Walnut Creek, CA 94598
sadoyle@lbl.gov

Full insight into the control of genomic sequences over many biological processes requires the analysis of the protein products. Only through the analysis of proteins on a genomic scale can we begin to understand the complexities encoded in the genome. Methods that allow for the production of proteins in a high-throughput manner are vital to achieve this goal. We have developed a system for high-throughput subcloning, protein expression and purification that is simple, fast and inexpensive. We utilize ligation-independent subcloning to create an expression vector encoding a N-terminal histidine tag. A dot blot expression screen was developed to analyze protein levels following expression in bacterial cultures, which facilitates the testing of multiple expression parameters if necessary. Protein purification in a 96-well format using Ni-NTA resin yields highly purified proteins. Using this system, we have optimized conditions to achieve a first pass rate success of up to 70% in prokaryotic systems, and are currently utilizing the expression screen to increase the efficiency of protein production from eukaryotic systems.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, the Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098, and the Los Alamos National Laboratory under contract No. W-7405-ENG-36.

69. Visualization and Analysis of Protein DNA Complexes

William McLaughlin[1], Xiang-jun Lu[1], Susan Jones[2], Janet Thornton[2], and Helen M. Berman[1]
[1]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, 610 Taylor Road, Piscataway, NJ 08854-8087
[2]Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT England
mclaugwi@rutchem.rutgers.edu


1BER

Simple and quantitative rules were created that can be used to discern DNA-binding protein structures in the Protein Data Bank and the Nucleic Acid Database. The rules are based on conserved structural characteristics analyzed by machine learning techniques. Where possible, a functional role has been assigned to each of the structural characteristics found.

We have also developed a computer program that depicts the interactions between DNA and proteins. This application creates schematic diagrams such as those seen in Figure 4 of Jones et al. (*J. Mol. Biol.* v287, pp877-896, 1999).

This work is funded by the Department of Energy (DE-FG02-96ER62166).

70. Structure/Function Analysis of Protein/Protein Interactions and Role of Dynamic Motions in Mercuric Ion Reductase

Susan M. Miller[1], Aiping Dong[2], Emil Pai[2], Matthew J. Falkowski[1], Richard Ledwidge[1], Anne O. Summers[3], and Jane Zelikova[3]
[1]Department of Pharmaceutical Chemistry, University of California - San Francisco
[2]Departments of Biochemistry, Medical Biophysics, Molecular and Medical Genetics, University of Toronto
[3]Department of Microbiology, University of Georgia, Athens
smiller@cgl.ucsf.edu

Mercuric ion reductase (MerA) is the key enzyme involved in bacterial pathways for detoxification of Hg(II) and organomercurials that result in the two-electron reduction of Hg(II) to elemental mercury. Extensive studies of this metal ion reductase have advanced our knowledge to the stage where detailed structure/function questions can be asked to gain deeper insight into how the structural components of the protein contribute to the efficient handling of the toxic metal ion. As these pathways are being incorporated into radiation resistant and other durable species for bioremediation purposes, these insights may prove invaluable for enhancing the activity of the protein and the whole pathway in the alternative organisms. In addition, with further insight into what features of the protein are critical for handling one metal ion, a second goal is to incorporate alternative ligands and properties to allow the protein to bind and reduce other toxic metals. Sequences of MerAs indicate the conservation of a multidomain catalytic core, in which a four-cysteine ligand exchange pathway for binding and reduction of Hg(II) has been identified. Two of the cysteines are found on the C-terminal segment of the protein that evidence suggests may require mobility for efficient catalysis. As one aspect of our studies, we are evaluating thermodynamic and kinetic properties of wild type and mutant MerAs with site-directed mutations in the ligand binding pathway and site of reduction, along with crystal structure analysis in order to evaluate the significance of mobility and other physicochemical properties on the efficiency of catalysis. In addition to the catalytic core, all but one reported MerA sequence also contain 1 or 2 N-terminal repeats of a domain (NmerA) with a conserved GMTCXXC metal-binding motif, the function of which has yet to be determined. A second aspect of our studies involves characterization of the NmerA structure, metal-binding properties and interactions with the catalytic core. To facilitate these studies, we have cloned and expressed the catalytic core and its single NmerA domain as separate proteins. As a third aspect of our studies, we are evaluating the effectiveness of the separate domains as participants in the mercury resistance operon in vivo. Results to date of these ongoing studies will be presented.

71. Investigating Protein Complexes by Crosslinking and Mass Spectrometry

Gregory B. Hurst[1], Robert L. Hettich[1], James L. Stephenson[1], Phillip F. Britt[1], Matthew Sega[3], Jana Lewis[1], Patricia K. Lankford[2], Michelle V. Buchanan[1], Edward C. Uberbacher [2], Ying Xu [2], Dong Xu[2], Jane Razumovskaya[3], and Victor N. Olman[2]
[1]Chemical Sciences Division and [2]Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge TN
[3]Genome Science and Technology Graduate School, University of Tennessee-Knoxville and Oak Ridge National Laboratory, Knoxville, TN
hurstgb@ornl.gov

In response to the Genomes to Life (GTL) Initiative, one component of the proposed Protein Complex Factory (PCF) at ORNL is a mass-spectrometry-based capability for high-throughput identification of protein complexes. The proposed strategy includes chemical crosslinking of interacting protein pairs, enzymatic digestion, affinity purification, and mass spectrometric analysis of the resulting crosslinked peptides to yield information on interacting protein pairs. Progress has been achieved on several elements of the proposed strategy. A special biotinylated family of crosslinkers will allow affinity isolation of either the crosslinked proteins, or of protolytic peptides from these proteins. A simple method for preparing these biotinylated crosslinking reagents has been tested. High

throughput will be achieved by performing crosslinking reactions on cell lysates, or fractions thereof, using a large array of reaction conditions. Some elements of this array of reaction conditions will yield crosslinking of a small subset of the protein complexes in the lysate or fraction. Crosslinking reactions have been performed under a variety of conditions in a 96-well format using model proteins. Identifying interacting proteins from their crosslinked peptides will require tandem mass spectrometry to yield partial amino acid sequence information from each of the two crosslinked peptides of the pair. Tandem mass spectrometry (MS-MS and MS-MS-MS) of a crosslinked peptide pair from bovine ribonuclease A shows a fragmentation pattern that allows confirmation of the identities of the two crosslinked peptides. Computational methods for interpreting the resulting mass spectra are under development.

72. High-Density Protein Microarrays

Judith Maples, Joseph Spangler, Yanhong Wang, and **Rajan Kumar**
Genome Data Systems, Inc.
rkumar@genomedatasystems.com

There is an increasing interest analysis of whole proteome analysis using high-density microarrays of proteins. Protein microarrays would form the basis of new diagnostics and research tools in the future. For diagnostic applications, protein microarrays can rapidly detect the presence or absence of biomarkers associated with particular diseases. In genomic and proteomic research, arrays of antibodies have been used to investigate how much of a given protein is expressed at a given time and place. However, the difficulties associated with protein microarrays are more difficult to address than DNA microarrays. Since the proteins are larger than DNA molecules, the individual protein molecules have to be deposited further apart resulting in lower sensitivity. Cross-reactivity of proteins is a major concern for protein microarrays. Genome Data Systems, Inc. has developed an innovative and highly flexible technology, called GeneCube, for fabrication, use and analysis of three-dimensional protein microarrays. The method allows mass-production of microarrays as well as stringent quality control. It also provides better accuracy and precision in comparison with conventional microarrays. The detection of signal from the microarrays is performed using a proprietary detection approach that extracts the signal from individual elements of the array without cross talk. During the current program, GeneCube microarrays were used to investigate interactions between proteins and antibodies, and to perform functional screening for potential substrates.

73. Advantages of Multi Photon Detectors in Protein Quantitation

A.K. Drukier
BioTraces Inc.
akd@biotraces.com

We are developing an integrated proteins detection system that is at least a hundred times more sensitive than current techniques. This protein quantification system capitalizes on multiphoton detector's (MPD) exquisite instrumental sensitivity to enable the highest sensitivity detection and high throughput. The ultra high sensitivity and very large dynamic range combine to make MPD instruments far superior to other methods of protein detection. Because of its ability to specifically detect co-resident labels, MPD technology in combination with prior-art protein microarrays (P-chips) permits about hundred-fold sensitivity improvement, mostly through new methods of non-specific biological background rejection.

MPD techniques: MPD is a proprietary detection system for the measurement of ultra-low amounts of selected radioisotopes ***[see www.biotraces.com]***. MPD enhanced biomedical methods have several advantages over existing methods: 1,000-fold improvement in sensitivity, enabling measurement of previously undetectable amounts of target substances; high dynamic range (8-9 decades), eliminating the need for sample concentration or dilution; use of extremely low levels of radioisotope, avoiding the classification of test samples as radioactive; cost savings due to decreased amounts of reagents and time for testing. With sensitivity better than a thousand atoms of ^{125}I, MPD marks a new milestone in detection where quantitation of sub-zeptomole amounts of biomaterial is possible. MPD techniques require less than one pCi of

isotope, which is about a 100-times less activity than in a glass of water.

MPD enhanced immunoassays: A new, super sensitive immunoassay (IA/MPD) that provides quantitative measurement of biological substances at levels as low as a femtogram/ml, *i.e.* sub-attomole sensitivities of IA/MPDs for several cytokines as well as the HIV-1 p24 antigen. Pilot studies compared the IA/MPD to prior art immunoassay methods. The unprecedented sensitivity of a family of IA/MPDs for interleukins (IL-1 beta, IL-4, IL-6, IL-10, IL-11, IL-12), as well as, the HIV-1 p24 antigen has been documented. This quantitatively accurate MPD immunoassays have a landmark sensitivity of about 1 fg/ml, *i.e.* better than 0.1 attomole/ml using a total specific activity that is below the natural radioactive background. Essential to the success of each IA/MPD has been our work on developing protocols and proprietary reagents for the reduction of nonspecific biological binding.

P-chips/MPD: We are extending IA/MPD to creation of supersensitive P-chips with MPD read-out (P-chip/MPD) targeting up to 256 different proteins. The pattern of activities is measured by the MPD-Imager with sensitivity better than 10 zeptomole/pixel. The sensitivity of such P-chip/MPD is clearly limited by the non-specific biological backgrounds. Preliminary results suggest that a sensitivity of 10-50 fg/ml can be achieved; this result has been achieved when quantitating several cytokines concurrently.

Detection of BW agents: MPD technology is applicable to the full spectrum of BW agents, including viruses, bacteria, algae and biotoxins. Initially, we propose the use of MPD enhanced detection methods for individual targets. We expect to complete the development of supersensitive P-chip/MPD for BW agents within a year. This universal system, able to detect all major groups of biological warfare agents, represents the most sensitive and pragmatic solution to the detection of BW threats. The proposed applications of MPD for the detection of biological warfare agents can be divided into two synergistic projects: use of a panel of MPD enhanced immunoassays for detection of biotoxins, and development of a universal biological warfare agent detector using supersensitive P-chips/MPD.

Bioinformatics

74. Understanding Protein Interactions

Xiaoqun Joyce Duan, Ioannis Xenarios, and David Eisenberg
UCLA-DOE Laboratory of Structural Biology and Molecular Medicine University of California, Los Angeles P. O. Box 951570, Los Angeles, California 90095-1570
joyce@mbi.ucla.edu

Networks of protein interactions control the lives of cells. One research interest in our lab focuses on understanding protein interactions and protein function using bioinformatic approaches. We have summarized studies from the scientific literature of interacting proteins in a database, the Database of Interacting Proteins (http://dip.doe-mbi.ucla.edu). DIP is designed to capture the layered information about protein interactions, which can be termed physical interactions and biological interactions. Biological protein interactions differ from the more general set of physical interactions in their prerequisite for specific protein states and the resultant transitions in the protein states of one or both of the interacting proteins. DIP contains information on physical interactions, including identities of the interacting proteins, their interacting regions, the binding affinity, and the experimental methods. LiveDIP, an extension of DIP, contains data on biological interactions, which are described in terms of protein states and state transitions. This data scheme provide a more complete picture of protein interactions inside cells. We developed advanced search tools such as Pathfinder and Batch searches to assemble pathways from currently available knowledge of protein interactions collated in LiveDIP. JDIP2D, is also developed to interactively explore interaction networks. It provides means of customized graph rendering, annotation, local storage and printing of the protein interaction networks. These data and tools of DIP offered some insights into protein interaction networks. Analysis of all the interactions in DIP indicates that many proteins form a single connected network of interactions accompanied by several smaller networks. An example of the pathway analysis tools applied to analyzing the pheromone response pathway in yeast suggests that the pathway functions in the context of a complex protein-protein interaction network and both positive and negative regulation are important in modulating signal intensity. Integrating gene expression data with this interaction network suggests some regulation mechanisms for signaling processes. Computational methods have also been developed to evaluate the overall quality of large-scale yeast two-hybrid experiments using gene-expression data. Future directions include expanding the database, providing additional tools for analyzing the validity of interactions, developing computational methods for predicting protein-protein interactions, and studying cell signaling on the genome scale.

References

1. Xenarios I, Fernandez E, Salwinski L, Duan XJ, Thompson MJ, Marcotte EM, Eisenberg D, DIP: the Database of Interacting Proteins: 2001 update. *Nucleic Acid Res* 2001 **29**(1): 239-241.
2. Duan XJ, Xenarios I, and Eisenberg D, Describing Biological Protein Interactions in Terms of Protein States and State Transitions: the LiveDIP Database. Manuscript in submission.

75. Automatic Discovery of Sub-Molecular Sequence Domains in Multi-Aligned Sequences: A Dynamic Programming Algorithm for Multiple Alignment Segmentation

Eric Poe Xing[1], Denise M. Wolf[1], Inna Dubchak[1], Sylvia Spengler[1], **Manfred Zorn**[1], Ilya Muchnik[2], and Casimir Kulikowski[2]
[1]Center for Bioinformatics and Computational Genomics, NERSC, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 U.S.A.
[2]Department of Computer Science and DIMACS, Rutgers University, Piscataway, NJ 08855 U.S.A.
mdzorn@lbl.gov

Automatic identification of sub-structures in multi-aligned sequences is of great importance for effective and objective structural/functional domain annotation, phylogenetic treeing and other molecular analyses. We present a segmentation algorithm that optimally partitions a given multi-alignment into a set of potentially biologically significant blocks, or segments. This algorithm applies dynamic programming and progressive optimization to the statistical profile of a multi-alignment in order to optimally demarcate relatively homogenous sub-regions. Using this algorithm, a large multi-alignment of eukaryotic 16S rRNA was analyzed. Three types of sequence patterns were identified automatically and efficiently: shared conserved domain; shared variable motif; and rare signature sequence. Results were consistent with the patterns identified through independent phylogenetic and structural approaches. This algorithm facilitates the automation of sequence-based molecular structural and evolutionary analyses through statistical modeling and high performance computation.

76. THE RDP-II (Ribosomal Database Project)

James R. Cole, Timothy G. Lilburn, Paul R. Saxman, Bonnie L. Maidak, Charles T. Parker, Sunandana Chandra, Ryan J. Farris, George M. Garrity, Thomas M. Schmidt, and James M. Tiedje
Center for Microbial Ecology, Michigan State University, East Lansing, MI 48824
colej@msu.edu

The Ribosomal Database Project - II (RDP-II) provides data, tools and services related to ribosomal RNA sequences to the research community. Through its website (http://rdp.cme.msu.edu), RDP-II offers aligned and annotated rRNA sequence data, analysis services, and phylogenetic inferences (trees) derived from these data. RDP-II release 8.1 (May 21, 2001) contains 16,277 prokaryotic, 5201 eukaryotic, and 1503 mitochondrial small subunit rRNA sequences in aligned and annotated format. Annotation goals include up to date name, strain and culture deposit information, sequence length and quality information, and references. In order to provide a phylogenetic context for the data, RDP-II makes available over 100 trees that span the phylogenetic breadth of life. Web based research tools are provided for comparing a user submitted sequence to the RDP-II database (Sequence Match), aligning a user sequence against the nearest RDP sequence (Sequence Align), examining probe and primer specificity (Probe Match), testing for chimeric sequences (Chimera Check), generating a similarity matrix (Distance Matrix), analyzing T-RFLP data (T-RFLP and TAP-TRFLP), and a java-based phylogenetic tree browser (Sub Trees). Release 8.1 debuted an updated sequence search and selection tool (Hierarchy Browser) and a new phylogenetic tree building and visualization tool built around the PHYLIP phylogeny inference package (Phylip Interface). In addition, release 8.1 includes an interactive tutorial to guide users through the basics of rRNA sequence analysis. This tutorial is suitable both for the researcher new to rRNA based phylogenetic analysis and as a teaching module for upper-level undergraduate and graduate classes. An ongoing effort at the RDP-II is the improvement of alignments in view of recent research on the ribosome and recent improvements in secondary structure based alignment algorithms. RDP-II is also working on methods to incorporate higher taxonomic information in its data. We expect these efforts to result in more accurate and timely data, and to increase the utility of RDP-II for the research community. The RDP-II email address for questions or comments is rdpstaff@msu.edu.

77. A Random Walk Down the Genomes: a Case Study of DNA Evolution in VALIS

Yi Zhou[1], Archisman Rudra[2], Salvatore Paxia[2], and Bud Mishra[2]
[1]Department of Biology, New York University
[2]New York University Courant Bioinformatics Group
yz237@nyu.edu

Modern biology is driven by large scale processing of heterogeneous data, which may come from diverse sources. This could be anything from a Genbank sequence to the result of some microarray experiment. The interfaces which let one access these different sources vary widely, so much so that a biologist needs to be an expert in very different areas of computer science: databases, networking, languages etc. Furthermore, the algorithms used to extract biologically significant information tend to be developed in an ad hoc manner. This leads to

very little code sharing between the data analysis algorithms with the concomitant increase in code complexity.

Instead of developing each tool ab initio, our bioinformatics system VALIS defines low level building blocks and uniform APIs which lets one use these from high level scripting languages. This enables biologists to write very simple scripts to perform fairly involved bioinformatics processing in a flexible fashion.

As an example we use the VALIS system to investigate the consequences of various cellular events on genomic DNA sequence evolution. How genomes evolve is a very important problem in biology. It will lead to better understanding on the mechanisms of cancer development, and more accurate analyses of phylogeny data.

We approach the study of sequence evolution by looking at statistical properties of the DNA sequences. In particular, we measure the long-range correlation properties of DNA sequences. Our approach is to estimate a few of these statistical parameters in the hope of distinguishing between different models of DNA evolution in coding and non-coding regions.

In order to study the scale-invariant long-range correlation of the DNA sequences, we view the DNA sequences as being generated from a random walk model. We first map the whole genomic DNA sequences following purine-pyrimidine binary rule: change purines (A/G) to +1 and pyrimidines (C/T) to -1. This creates a 'DNA walk' along the genome. The 'DNA walker' moves either up or down at every base pair according to the binary map of the DNA sequence. If there is no long-range correlation, the walk is a realization of a Brownian motion. Otherwise, we observe a 'walker' with long-term memory and thus a Fractional Brownian motion. Those two processes can be characterized by different values of the Hurst exponent (H). H=0.5 for Brownian motion and H>0.5 for Fractional Brownian motion, i.e. higher H values suggests the presence of stronger long-range correlation. We use many different methods to estimate H, for example, R/S analysis and detrended fluctuation analysis (DFA).

We have analyzed various genomes using VALIS: bacteria, invertebrate and vertebrate. We observe a consistent difference in H in the coding regions compared to the non-coding regions. The H values tend to be higher in the non-coding regions than in the coding regions. Thus, the DNA walk down the bacterial coding region sequences behaves as a Brownian motion ($H \sim 0.5$), while it acts as a Fractional Brownian motion in the non-coding regions ($H>0.5$). For other organisms, such as yeast, the difference persists: yeast has $H \sim 0.54$ in the coding regions, versus $H \sim 0.61$ in the non-coding regions. The higher H values in non-coding regions indicate that the sequences in the non-coding regions possess much stronger long-range correlation than those in the coding regions. In addition, the H values in different regions increase with the evolutionary position of the corresponding organism. This suggests that there are some cellular events that tend to make DNA sequences more correlated as evolution proceeds.

Based on our observations, we hypothesize that the differences in the strengths of long-range correlation in DNA sequences are caused by the counteraction of two sets of biological events. One set includes insertion, deletion events caused by DNA polymerase stuttering and transposons, which tend to increase DNA long-range correlation. And the other set includes natural selection and DNA repair mechanisms, which try to eliminate the long-range correlation caused by the former events. However, the coding regions are under a higher natural selection pressure and possess the transcription-coupled DNA repair mechanism that is unique to them. Thus, the stronger correlation-elimination forces in the coding regions can explain the weaker long-range correlation observed there than that in the non-coding regions. And the higher flexibility offered by larger genome sizes in the higher organisms allows the increase of long-range correlation in DNA sequences along the evolution tree.

To test our hypothesis, we designed a 'Genome Grammar'. This is a stochastic grammar with primitives for many kinds of mathematical probability distributions. We can even generate a sequence with the same probability distribution as measured from biological data. Furthermore, there are tools that let one apply some hypothesized

processes act on sequences obtained from the grammar. This enables biologists to apply any model and conduct evolutionary experiments 'in silico'.

Our observations also have potential significance for biotechnology application. Taking the advantage of highly efficient statistical algorithms in VALIS, the discovery of statistical differences in DNA coding and non-coding regions may lead to potential in vitro biochemistry technologies that can efficiently detect coding and non-coding regions without the effort of DNA sequencing.

78. A Graph Data Model to Unify Biological Data

Frank Olken
Lawrence Berkeley National Laboratory
Olken@lbl.gov

Federated (or mediated) database systems typically require that we describe each participating database's schema in a common data model, e.g., relational. Such a common data model facilitates the construction of queries which span the various databases (and types of data).

In this work, we suggest that a graph data model, i.e., labeled graphs, either directed or undirected, could better serve as this common data model for biological data. We note the ubiquity of graphs in biological datasets: taxonomies, phylogenetic trees, metabolic networks, signaling networks, genetic regulatory networks, chemical structure graphs, contact graphs, partial orders in genetic mapping, overlap graphs in physical mapping and shotgun sequence assembly, DNA sequences as linear graphs, etc. Finally, we review graph data modeling and query language efforts in the database community and point out open problems in graph data modeling, query language design, query complexity for data management in biology.

79. Protein Data Bank: Unifying the Archive

Gary Gilliland and The PDB Team
Research Collaboratory for Structural Bioinformatics Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, 610 Taylor Road, Piscataway, NJ 08854-8087; National Institute of Standards and Technology, Biotechnology Division and Informatics, Data Center, 100 Bureau Drive, Gaithersburg, MD 20899-8314; and San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0537
gary.gilliland@nist.gov

The Protein Data Bank (PDB; http://www.pdb.org/) is the single worldwide archive of structural data of biological macromolecules. All data in the archive have been validated, and a uniform archive has been released for the community. A collection of mmCIF data files for the PDB archive has been made available at ftp://beta.rcsb.org/pub/pdb/uniformity/data/mmCIF/.

A utility application that converts the mmCIF data files to the PDB format has also been released to provide support for existing software.

The Protein Data Bank is operated by the Research Collaboratory for Structural Bioinformatics (RCSB) and is supported by funds from the National Science Foundation, the Department of Energy, and two units of the National Institutes of Health: the National Institute of General Medical Sciences and the National Library of Medicine.

80. Protein Structure Predictions by PROSPECT

Dong Xu, Dongsup Kim, Christal Secrest, Victor Olman, and **Ying Xu**
Protein Informatics Group, Life Sciences Division, Oak Ridge National Laboratory
xyn@ornl.gov

Protein threading represents one of the key computational techniques for protein fold recognition and protein backbone structure

prediction. We have previously developed a computer software PROSPECT, using protein threading as its core technology. PROSPECT employs a divide-and-conquer algorithm for finding the globally optimal alignment between a query sequence and a template structure. Significant improvements and additions have been made to the PROSPECT system in the past year, which can be summarized as follows.

1. We have developed a capability for assessing the reliability of PROSPECT's fold recognition prediction. It is well-known that there is no theoretically sound way, yet, to normalize threading scores across all queries/templates with different sequence lengths, different amino acid compositions, and different geometric and physical features, making assessing threading scores difficult. By threading each template structure in our template database (with over 2000 structures) against all sequences in the FSSP database, we can get a threading score distribution for the template against all FSSP proteins. We have trained a neural network to map each query-template threading score along with the template's score-distribution and various other compositional, geometric and physical parameters of the template-query pair to a real value in [0,1], such that this value reflects the percentage of structurally-alignable positions between the query and the template (with 1 representing two structures 100% alignable; and 0 for no significant structure alignment). This mapping provides a highly useful measure for PROSPECT's predictions.
2. We have improved PROSPECT's threading energy function by including family-specific profiles and profile-profile alignments, and by re-parameterizing our current energy terms through better statistical treatments of the structure database information and using significantly large data set. This has resulted in a 10+% increase in PROSPECT's prediction accuracy on a large test set, compared to the previous version of PROSPECT.
3. We have developed a capability for automatically decomposing a solved protein structure into protein domains. This is an essential step in automatically updating our structure template database. This unique capability solves the domain decomposition problem as a maximum flow problem. It can be potentially used by large protein-structure depositories like PDB for its automatic updates on protein domains.

We have recently applied PROSPECT to a number hypothetical proteins of *Shewanella oneidensis* MR-1, which were identified to be related to metal-reduction through microarray gene expression experiments and data analysis by J. Zhou's lab at ORNL. Detailed structures of four such proteins will be presented in this presentation.

References

1. Y. Xu, and D. Xu, "Protein Threading using PROSPECT: design and evaluation", *Protein: Structure, Function, Genetics*, **40**:343 - 354, 2000.
2. D. Xu, K. Baburaj, C. B. Peterson, and Y Xu, "A Model for the Three Dimensional Structure of Vitronectin: Predictions for the Multi-Domain Protein from Threading and Docking", *Proteins: Structure, Function, Genetics*, **44**:312-320, 2001.
3. D. Xu, O. Crawford, P. LoCascio, and Y. Xu, "Application of PROSPECT in CASP4: characterizing protein structures with new folds", *Proteins: Structure, Function, Genetics special issue on CASP4* (by invitation), 2001 (in press).
4. Y. Xu, D. Xu and V. Olman, "A practical method for interpretation of threading scores: an application of neural networks", *Statistica Sinica Special issue on Bioinformatics*, 2001 (in press).
5. D. Xu and Y. Xu, "Computational Studies of Protein Structure and Function using Threading Program PROSPECT", In *Protein Structure Prediction: Bioinformatic Approach* (eds by Igor Tsigelny), International University Line (IUL) Publishers (by invitation), 2001 (in press).
6. Y. Xu, D. Xu, and H. N. Gabow, "Protein Domain Decomposition using a Graph-Theoretic Approach", *Bioinformatics*, **16** (12), 1091 - 1104, 2000.

81. Protein Fold-Recognition Using HMMs and Secondary Structure Prediction

Kevin Karplus
University of California, Santa Cruz
karplus@soe.ucsc.edu

The protein-folding problem, in its purest form, is too difficult for us to solve in the next several years, but we need structure predictions now. One solution is to try to recognize the similarity between a target protein and one of the thousands of proteins whose structure has been determined experimentally. For very similar proteins, the relationships are easy to find and good models can be built by copying the backbone (and even some sidechains) for the homologous protein of known structure. For less similar proteins (in the "twilight zone"), the fold-recognition problem is more challenging, but it is often possible to find useful similarities.

Using evolutionary information helps enormously in recognizing remote relationships, and one convenient way to summarize a family of homologs is with a hidden Markov model (HMM). Homologs can be found and an HMM built by an iterated search, starting from a single target sequence. The resulting target HMM can be used to score the sequences of all proteins of known structure.

Similarly, homologs can be found and HMMs built for template proteins of known structure and used to score the target sequence. Combining both target-model and template-library results reduces the false positive rate.

Some further improvements can be made by predicting local structural properties of the target sequence (such as secondary structure or solvent accessibility) and adding these predictions to the HMM used to score the template sequences.

Fold-recognition techniques based on these HMMs have performed quite well in blind prediction experiments (CASP2, CASP3, and CASP4) and are doing better than threading techniques based on pairwise potentials.

82. Protein Engineering in Structural Genomics

Patrice Koehl and Michael Levitt
Department of Structural Biology, Fairchild Bldg, Stanford University, Stanford, CA 94305-5126
koehl@csb.stanford.edu

The emphasis of our project is placed on the design of novel proteins that may serve as catalysts needed for bioremediation. Our approach can be divided into two steps: identify target protein structures that would provide the desired functions, and search for sequences that make these protein structures both stable and unique. Our efforts have focused on the later, namely on automated protein sequence design. We have made significant progress in characterizing the sequence space compatible with a protein structure, and have shown that this information can prove valuable for protein structure prediction:

(1) Measuring the size of the sequence space compatible with a protein structure
It is well known that certain structures are more commonly observed among proteins than others. Highly designable structures are more likely to have been found through the process of evolution, since they are more robust to random mutations. We have developed a new approach to explore and quantify the sequence space associated with a given protein structure. We have shown that our measure of the protein sequence space compatible with a given fold correlates with the usage of the fold observed among naturally occurring sequences. Our results also suggest that the designability of a protein (i.e. the number of sequences possessing tbe structure of interest as their non-degenerate energy ground state) can be derived from the knowledge of its topology alone. As a consequence, we anticipate that our method for sequence space exploration will prove useful for identifying highly designable folds, which will represent attractive targets for protein design.

(2) Application of protein sequence design to protein structure prediction
The goal of the inverse protein folding problem is to identify amino acid sequences that stabilize a given target protein conformation. Methods that attempt to solve this problem have proved useful for protein sequence design. We have shown that the same methods can provide valuable information for protein fold recognition and for *ab initio* protein

structure prediction. We also derived a new measure of the compatibility of a test sequence with a target model structure, based on computational protein design. The protein structure is used as input to design a family of low free energy sequences, and these sequences are compared to the test sequence, using a metric in sequence space based on nearest neighbor connectivity. We have found that this new measure is powerful enough to recognize near-native protein structures among non-native models.

83. Classifying G-Protein Coupled Receptors with Support Vector Machines

Rachel Karchin[1], Kevin Karplus[2], and David Haussler[3]
[1]University of California, Santa Cruz, Computer Science
[2]University of California, Santa Cruz, Computer Engineering
[3]Howard Hughes Medical Institute
rachelk@soe.ucsc.edu

The enormous amount of protein sequence data uncovered by genome research has increased the demand for computer software that can automate the recognition of new proteins. We discuss the relative merits of various automated methods for recognizing G-protein coupled receptors (GPCRs), a superfamily of cell membrane proteins. GPCRs are found in a wide range of organisms and are central to a cellular signaling network that regulates many basic physiological processes. They are the focus of a significant amount of current pharmaceutical research because they play a key role in many diseases. However, their tertiary structures remain largely unsolved. The methods described in this paper use only primary sequence information to make their predictions. We compare a simple nearest neighbor approach (BLAST), methods based on multiple alignments generated by a statistical profile hidden Markov model, and methods, including support vector machines, that transform protein sequences into fixed-length feature vectors. The last is the most computationally expensive method, but our experiments show that, for those interested in annotation-quality classification, the results are worth the effort. In two-fold cross-validation experiments testing recognition of GPCR subfamilies that bind a specific ligand (such as a histamine molecule), the errors per sequence at the minimum error point (MEP) were 13.7% for multi-class SVMs, 17.1% for our SVMtree method of hierarchical multi-class SVM classification, 25.5% for BLAST, 30% for profile HMMs, and 49% for classification based on nearest neighbor feature vector (kernNN). The percentage of true positives recognized before the first false positive was 65% for both SVM methods, 13% for BLAST, 5% for profile HMMs and 4% for kernNN.

We have set up a web server for GPCR subfamily classification based on hierarchical multi-class SVMs at http://www.soe.ucsc.edu/ research/ compbio/gpcr-subclass. By scanning predicted peptides found in the human genome with the SVMtree server, we have identified a large number of genes that may encode GPCRs.

A list of our predictions for human GPCRs is available at http://www.soe.ucsc.edu/research/ compbio/gpcr_hg/class_results. We also provide suggested subfamily classification for 18 sequences previously identified as unclassified Class A (rhodopsin-like) GPCRs in GPCRDB, available at http://www.soe.ucsc.edu/ research/compbio/ gpcr/classA_unclassified/.

84. Protein Structure Determination Through Combining Protein Threading and Sparse NMR Data

Ying Xu, Dong Xu, Dongsup Kim, and Oakley Crawford
Protein Informatics Group, Life Sciences Division, Oak Ridge National Laboratory
xyn@ornl.gov

Protein structural information derived from protein threading and NMR experiments could complement each other. Fully utilizing the available information from the two sources could lead to solutions of protein structures neither one alone can solve. When applicable, protein threading can provide protein backbone structures with reasonable accuracies (e.g, 4-6 angstroms). It is estimated that threading methods could potentially be applicable to 60-70%

of all soluble proteins. NMR experiments typically apply to small proteins (< 30 KD). As target proteins become larger, the fraction of assignable NMR peaks drops significantly, which results in an insufficient amount of NMR restraints for accurate structure solution. We have developed a computational capability for protein structure solution through combining sparse NMR data and protein threading. It consists of two components: (a) protein fold recognition and backbone prediction by NMR data-constrained threading; and (b) NMR structure calculation by molecular dynamics and energy minimization, using a threaded structure as starting point. We have demonstrated that a small number of NMR restraints can significantly improve the prediction accuracy by threading, and that when starting from a predicted backbone structure (with accuracy about 4 angstrom), NMR structure calculation only requires a small fraction of NMR restraints typically needed to reach a certain level of accuracy. To make this computational capability practically useful, a capability for assigning (sparse) backbone NMR peaks is essential. We have developed a computational framework for assigning backbone NMR peaks. The framework models peak assignments as a constrained bipartite matching problem. While we demonstrated that a rigorous solution is highly challenging (we proved the problem is NP-hard), we have developed a rigorous and reasonably efficient algorithm by taking advantage of the discerning power of our assignment function. This framework is the first rigorous formulation, which is capable of incorporating all relevant information involved in peak assignments. Our preliminary assignment results are highly encouraging. Collaborations are currently under way to solve a number of large proteins using this computational framework, with NMR labs.

References:

1. Y. Xu, D. Xu, D. Kim, V. Olman, J. Razumovskaya, and T. Jiang, "Automated Assignment of Backbone NMR Peaks using Constrained Bipartite Matching", *IEEE Computing in Science and Engineering special issue on bioinformatics*, 2001 (in press).
2. Y. Xu and D. Xu, "Protein Structure Prediction by Protein Threading and Partial Experimental Data", in *Current Topics in Computational Molecular Biology* (Eds, Jiang, Xu, Zhang), MIT Press, 2001 (in press).
3. Y. Xu, D. Xu, O. Crawford, J. R. Einstein, "A computational method for NMR-constrained protein threading", *Journal of Computational Biology*, 7:449 - 467, 2000.
4. G. Lin, Z. Chen, T. Jiang, J. Wen, J. Xu, Y. Xu, "Approximation Algorithms for NMR Spectral Peak Assignment", 2001 (submitted).

85. GAP: Genomics Annotation Platform

Konstantin M. Skorodumov[1], Evgeny Raush[1], Maxim Totrov[1], Ruben Abagyan[2], and **Matthieu Schapira**[1]
[1]Molsoft LLC
[2]The Scripps Research Institute
matthieu@molsoft.com

The challenge of functional genomics—assigning functions to sequenced genes—is critical for the rapid evolution of modern medicine. Computational approaches can accelerate dramatically annotation efforts by producing predictive functions that can then be rapidly confirmed. Molsoft is building a genomics annotation platform, GAP, based on in-house computational biology and intranet software. Both comparative genomics and structural genomics tools are being implemented, and will allow rapid identification of predictive protein functions and protein-protein interactions. The system relies on a relational database infrastructure and provides online graphic user interface.

86. Genome to Proteome and Back Again: ProteomeWeb

Carol S. Giometti[1], Sandra L. Tollaksen[1], Gyorgy Babnigg[1], Tripti Khare[1], Claudia I. Reich[2], Gary J. Olsen[2], John R. Yates III[3], Jizhong Zhou[4], Ken Nealson[5], and Derek Lovley[5]
[1]Argonne National Laboratory, Argonne, IL
[2]University of Illinois, Urbana-Champaign, IL
[3]The Scripps Institute, La Jolla, CA
[4]Oak Ridge National Laboratory, Oak Ridge, TN
[5]University of Massachusetts, Amherst, MA
csgiometti@anl.gov

Complete genome sequences give rise to open reading frame (ORF) databases that can be translated into the theoretical amino acid sequences of all proteins predicted to be encoded. In the context of proteome analysis, these ORF databases are the foundation of protein identifications. Proteins from complex mixtures are digested using site-specific proteases and the masses of the peptides are compared with the hypothetical peptide masses from the protein sequences predicted by the ORFs. As part of the Department of Energy Microbial Genome Program, we are using two-dimensional gel electrophoresis coupled with tandem mass spectrometry to identify and quantify the proteins expressed by a variety of energy- and bioremediation-related microbes, including *Methanococcus jannaschii, Shewanella oneidensis,* and *Geobacter sulfurreducens*. Sufficient data has been assimilated to allow comparisons among microbial proteomes in the context of constitutive protein expression and the modulation of protein expression in response to specific environmental conditions. Data acquisition and management are handled by using the Oracle relational database software together with a World Wide Web interface. A public web site (ProteomeWeb; http://proteomes.pex.anl.gov) with customized tools is available to enable users to query the protein identifications and experimental data. This web site includes links to genome and protein sequence databases as well as metabolic pathway databases, providing an integrated environment for the interpretation of the proteome results. The proteome studies are adding value to existing genome sequence information, providing data on the relative abundance of different ORF products, the conditions of their expression, and post-translational processing. However, mechanisms for genome databases to access and utilize these proteome data still need to be developed.

This work is supported by the U. S. Department of Energy, Office of Biological and Environmental Research, through the Microbial Genome and NABIR Programs, under contract No. W-31-109-ENG-38.

87. Computational Experiments on RNA Phylogeny

Frank Olken[1], James R. Cole[2], Gary J. Olsen[3], Craig A. Stewart[4], David Hart[4], Donald K. Berry[4], and Sylvia J. Spengler[1]
[1]Lawrence Berkeley National Laboratory
[2]Michigan State University
[3]University of Illinois, Urbana Champaign
[4]Indiana University
olken@lbl.gov

This work describes computational experiments aimed at automated construction of high quality phylogenetic trees from RNA sequences taken from the Ribosomal Database Project (RDP). Due to computational constraints and limitations of earlier multiple sequence alignment codes current production operations at the Ribosomal Database Project use hand tuned multiple sequence alignments and trees constructed with the WEIGHBOR neighbor joining code (by W.J. Bruno, N.D. Socci, and A.L. Halpern) Here we describe efforts to construct maximum likelihood phylogenetic trees from the RDP data set using RNACAD (by M. Brown) to construct the multiple sequence alignments and a parallel version of the fastDNAml code (G. Olsen, parallelization by C. Stewart, D. Hart, and D. Berry). to construct the maximum likelihood phylogenetic trees.

88. Identifying Transcription Factor Binding Sites by Cross-Species Comparison

Lee Ann McCue, William Thompson, C. Steven Carmack, and Charles E. Lawrence
The Wadsworth Center, New York State Department of Health, Albany, NY
mccue@wadsworth.org

We have developed a phylogenetic footprinting method with the goal of identifying complete sets of transcription factor (TF) binding sites in bacterial genomes. This method employs an extended Gibbs sampling algorithm to identify sites by cross-species comparison. The *Escherichia coli* genome sequence and a database of experimentally verified regulatory sites were used to test this method. Using this data

and the genome sequence data from nine additional gamma proteobacterial species, we have evaluated our ability to predict TF binding sites, and addressed the questions of which species are most useful and how many genomes are sufficient for comparison with respect to phylogenetic footprinting. In a study set of 166 *E. coli* genes with experimentally identified TF binding sites upstream of the orf, we found that orthologous promoter data from just 3 additional species were sufficient for ~80% of predicted sites to correspond to experimentally reported sites. Also, the species characteristics that most influenced our results were phylogenetic distance, genome size, and natural habitat. We performed simulations using randomized data to determine the critical values for statistical significance of our predictions ($p = 0.05$). We found that the inclusion of a very closely related species (*Salmonella typhi*) was beneficial despite substantially increasing the critical value. We are applying this technology to the genomes of microbes that are of environmental interest and for which there is little knowledge of transcription regulation. Preliminary results for *Synechocystis* PCC6803 will be presented.

89. VISTA: Integrated Tool for Comparative Genomics

I. Dubchak[1], Lior Pachter[2], A. Poliakov[1], I. Ovcharenko[1], and E. Rubin[1]
[1]Lawrence Berkeley National Laboratory
[2]University of California Berkeley
ildubchak@lbl.gov

One of the more powerful algorithms available to identify functional regions in genomic DNA (this includes both genes and surrounding regulatory elements) involves comparative sequence analysis. They have proved to be especially efficient in finding and analyzing conserved non-coding elements potentially playing a role in gene regulation. The deluge of genomic sequence that is rapidly appearing in databases is leading to the need for faster and more robust programs for analyzing the data. For example the practice of aligning single genes only a few kilobases long has been replaced by the need to align hundreds of kilobases of BACs or even entire genomes. The algorithmic challenges posed by these large datasets have been accompanied by user interface challenges, such as how to visualize information related to enormous datasets and how to enable users to interact with the data and the processing programs.

We have developed an integrated tool incorporating a novel alignment program for large DNA sequences AVID (Bray and Pachter, in preparation) and an associated visualization tool VISTA (Mayor et al., 2000), which serves as the platform for large-scale comparative analysis of genomic sequences. Its visual output is clean and simple, allowing the user to easily identify conserved regions. Similarity scores are displayed for the entire sequence, thus helping in the identification of shorter conserved regions, or regions with gaps.

There are various modifications of VISTA for solving particular biological problems: cVISTA (complementary VISTA) is used for looking at differences between recently evolved species such as comparing mice to rats or humans to chimpanzees; multiVISTA allows to visualize several related alignments on the same scale. When orthologous sequences of three species are available, we can also apply a statistical method for calculating cutoffs to define noncoding sequences that are conserved because of functional constraints (Dubchak et al., 2000). VISTA has been implemented as a platform-independent stand-alone software package written in Java and as a Web server located at http://www-gsd.lbl.gov/vista. It is extensively used in Genome Science Department for mouse-human comparative studies and has become the main comparative sequence analysis tool of several large sequence generation centers.

1. Mayor C., Brudno M., Schwartz J. R., Poliakov A., Rubin E. M., Frazer K. A., Pachter L. S. and Dubchak I. (2000) VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, **16**, 1046-1047.
2. Dubchak I., Brudno M., Loots G. G., Mayor C., Pachter L., Rubin E. M. and Frazer K. A. (2000) Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Research*, **10**.

90. Beyond Terascale Biological Computing: GIST and Genomes To Life

Philip LoCascio, Doug Hyatt, Frank Larimer, Manesh Shah, Inna Vokler, and Ed Uberbacher
Computational Biology Program, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee
http://compbio.ornl.gov/
locasciop@ornl.gov

High performance computing played a critical part in the successful completion of the working draft of the human genome, and continues to be necessary to handle the ever-increasing flood of new biological data. We have successfully met the first rounds of such computing for biology with the Genomic Integrated Supercomputing Toolkit (GIST). GIST provides a transparent, fault tolerant interface for the research community to an ever increasing suite of accelerated massively parallel biological applications. It also demonstrates key concepts that can be extended for large-scale computation for the Genomes to Life (GTL) program. GIST and associated data sets are accessible via the WWW interfaces of the Genome Analysis Toolkit, Genome Channel and other ORNL tools, and optionally a command line interface for biologists developing new applications.

With the advent of the DOE Genomes To Life Program, we are actively developing new technologies that will help biologists link large-scale experimentally derived biological data with increasingly sophisticated computational analyses. Our basic approach is to support the mode of operation where computational biology is concerned with transactional tool usage upon data and the construction of "recursive" pipelines of analyses, which ultimately execute on supercomputer resources (via GIST). The central theme here is to organize the libraries of biological tools around the available biological data types, and use the existing methodology of context dependent XML schema to classify both the data types and the biological tools. In this way, it becomes possible for users to (i) automatically detect which tools and which data can be combined in a valid operation, (ii) configure linked sets of analysis steps without detailed knowledge of computing or tools, and (iii) record transactions in an RDBMS to look for dependencies and redundancies. Furthermore, existing user interfaces to biological tools in any browser format can be instantly coupled with the appropriate tool-data combinations and linked transparently to high performance operations.

Where available, existing biomedical community templates are being used as the basis for standards which could be established across the GTL enterprise. With new data types rapidly becoming available, a strategy that reuses existing tools and interfaces, supports new tools, and makes it possible to combine tools to perform novel analyses, will be the most effective way to manage software complexity and development cost. Additionally, where data types are not "naturally" aligned, it will be possible to create "filters" for the most common types of conversions. e.g. FASTA → Masked FASTA.

This intermediate software layer can serve as an effective conduit between users, with their novel complex biological data sets, and the emerging beyond terascale computing infrastructure. Using this approach it should be possible to create friendly interfaces to new classes of algorithms that are built of both new and old components, with the flexibility necessary to tackle biological problems of increasing complexity. Libraries of software for tasks such as metabolic pathway reconstruction, gene regulatory network modeling and cell modeling can be constructed, supported and utilized by the community if organized and accessed in thus manner.

(Research sponsored by the Office of Biological and Environmental Research, US DOE under contract number DE-AC05-00OR22725 with UT-Battelle, LLC)

91. A Computational Pipeline for Genome-Scale Analysis of Protein Structures and Functions

Serguei Passovets, Manesh Shah, Li Wang, Dong Xu, and **Ying Xu**
Life Sciences Division, Oak Ridge National Laboratory
xyn@ornl.gov

We are expecting over 1000 genomes will be sequenced within the next 5-10 years. A significant percentage of the genes, to be identified computationally in these genomes, will have not have known functions, which are detectable by sequence-based homology search tools like PSI-BLAST. In our recent experience in CASP4, we found that threading-based protein fold recognition tools, like PROSPECT, can clearly detect more remote homologs than sequence-based methods can. We have recently developed a computational pipeline for automated protein structure predictions. The main components of the pipeline are: (a) a toolkit consisting of essential protein analysis tools, (b) a client/server system which provides access to the tools, (c) a pipeline manager which coordinates the processing tasks for a given analysis request, and (d) a web interface for query submission.

The pipeline operations can be categorized into three distinct phases: 1) protein triage, 2) threading-based structure prediction and 3) sequence based function determination. Protein triage phase uses PRODOM (for domain parsing), SOSUI (for classification into globular or membrane protein), SignalP (for identifying signal peptide cleavage sites) and PSI-BLAST (for sequence homology in PDB, Swissprot and other databases). Structure prediction phase uses SSP (a secondary structure prediction tool developed by our group), PROSPECT (for protein fold recognition), MODELLER (for atomic model construction) and WHATIF (for structure quality assessment). Sequence based function determination phase (not yet implemented) will use protein family classification tools Pfam, Motif and PRINTS. The pipeline manager invokes different tools depending on the user input and logic of the prediction process and controls the data and analysis flow of the pipeline. XML technology is used for data exchange between the web interface, the pipeline manager and the tools.

Initial applications of the pipeline will be done on proteins of the cyanobacteria genomes, currently being annotated at ORNL.

Reference:

1. D. Xu, O. Crawford, P. LoCascio, and Y. Xu, "Application of PROSPECT in CASP4: characterizing protein structures with new folds", *Proteins: Structure, Function, Genetics special issue on CASP4* (by invitation), 2001 (in press).

92. WIT3 – A New Generation of Integrated Systems for High-Throughput Genetic Sequence Analysis and Metabolic Reconstructions

N. Maltsev, G. X. Yu, E. Marland, S. Bhatnagar, R. Lusk, and E. Selkov
Mathematics and Computer Science Division
Argonne National Laboratory
maltsev@mcs.anl.gov

During the past decade, the scientific community has witnessed an unprecedented accumulation of gene sequence data and data related to the physiology and biochemistry of organisms. Availability of such information allows moving to the next level of understanding of life processes by shifting of focus of investigation from individual genes and proteins to understanding of functionality of the biological systems as a whole. Therefore development of integrated computational environments that provide access to genomic data, information describing metabolic and regulatory networks, as well as computational tools for navigation, comparisons, cross-correlations and analysis of this data is critical for further advancement of biological science.

During the past years the Computational Biology group at Argonne designed and implemented a family of systems for genetic sequence analysis and metabolic reconstructions of newly sequenced genomes. The WIT2 system developed at Argonne National Laboratory (http://wit.mcs.anl.gov/WIT2) is an interactive integrated information system that is used by the scientific community worldwide for comparative analysis of the genomes and metabolic reconstructions from the sequence data. However,

advances in software development and in computational capabilities now offer the opportunity to significantly enhance capability of the WIT2 system to perform high-throughput analysis of microbial genomes at a rate compatible with the increased rate of genome sequence completion.

Computational biology group at ANL is now developing the next generation of such systems, WIT3. WIT3 now contains analysis of 74 completely sequenced microbial genomes and has the following new features that allow improving performance and enhancing genome analysis:

a) WIT3 is based on relational database (Oracle). Such systems architecture increases overall performance and robustness of the system, simplifies additions of the new genomes and updates procedures.
b) Combination of SQL and Perl-based search engines allows fast execution of complex queries
c) Representation of the data in a structured XML format simplifies analysis, representation, visualization and exchange of the data.

We have also developed a number of new tools and algorithms for automated analysis of the genomes. Such clustering algorithms allow integrating the results of analysis of genomic data by a variety of comprehensive bioinformatics tools (e.g. InterPro, Blocks, CATH) and increasing sensitivity and reliability of the genetic sequence analysis. We are developing tools to support comparative analysis of the metabolic and regulatory networks, and user-driven as well as automated metabolic reconstructions from the sequence data. During the past months we have developed a new user interface that allows extensive visualization of the sequence data (e.g. domain organization, genetic maps, conserved chromosomal gene clusters, phylogenetic trees). New tools for visualization and navigation of the metabolic networks are also being developed.

Data processing is done using scalable computational supercomputing resources available at MCS.

93. Comparative and Collaborative Bioinformatics Systems to Promote Mammalian Phenotype Analysis and the Elucidation of Regulatory Networks

Erich Baker[1,5,6,7], Doug Hyatt[1,5], Barbara Jackson[1,2,6,7], Gwo-Liang Chen[1,6,7], Denise Schmoyer[1,2,6], Yesim Aydin-Son[1,5], David McWilliams[1,5], Fred Baes[1,6,7], Stefan Kirov[1,5], Michael Galloway[1,6], Michael Leuze[2], Line Pouchard[2], Brynn Jones[1], Ed Michaud[1,5], Bem Culiat[1,5], Gene Rinchik[1,5,7], Dabney Johnson[1,5,7], Ed Uberbacher[1,5], Darla Miller[1,7], **Frank Larimer**[1,5], Jay Snoddy[1,4,5,6,7], ORNL Life Sciences Division, and the Tennessee Mouse Genome Consortium
[1]Life Sciences Division, Oak Ridge National Laboratory
[2]Computer Science and Mathematics Division, Oak Ridge National Laboratory
[3]University of Tennessee Health Science Center
[4]University of Tennessee Center of Excellence in Genomics and Bioinformatics
[5]UT-ORNL Graduate School in Genome Science and Technology
[6]Tennessee Mouse Genome Consortium—Bioinformatics Core
[7]Tennessee Mouse Genome Consortium—Neurophenotyping project, PI: Dan Goldowitz UTHSC
fwl@ornl.gov

Phenotypes are complex, highly diverse across species, and variable within populations of a species. We will soon have complete lists of the genes in the mouse, human, and a few other metazoan species, but we still do not understand how these genes act together to create phenotypes from genotypes. Gene regulatory networks (GRNs) are thought to be a major component in "reading out" genotypes and creating phenotypic complexity, diversity, and variability. Analysis of the networks that create phenotypes will require new data and approaches. Collaborations under way at Oak Ridge National Laboratory (ORNL) and the Tennessee Mouse Genome Consortium (TMGC) (www.tnmouse.org) are good examples of what is needed. These studies will also require developing new aspects of collaborative and comparative bioinformatics. Our group, working with others, will need to create a

Semantic Web for phenotype and GRN analysis. This semantic web should, like the current web, allow scientists to directly share and analyze data that is placed on the web; more importantly, data for a semantic web is organized so that computers can also directly analyze large data sets placed on the web. This computer-based analysis should create useful inferences and assist users in datamining.

Information systems developed to directly support the sociological aspects of web-based collaborations include a ***TMGC member database (ver. 1.0)***, ***TMGC protocol-expertise database (ver. 0.1)***, and a collaborative ***TMGC Web Content Management System (ver. 1.0)***. Other bioinformation systems discussed below are needed to support four different, but interrelated, research approaches. The first two approaches include alternative ways to find and analyze new allelic variants that have a phenotypic consequence. The third approach will start from the molecular phenotypes of gene expression (e.g. RNA expression arrays or proteomes). The final interrelated approach will start from a comparative analysis of genome sequences in the chordates to lay the groundwork to understand the evolution of genomes and gene regulatory networks.

Within a species, genotypic variation—either created experimentally or occurring naturally—may cause system perturbations that result in observed variation in phenotypes. Our group is developing bioinformation systems to assist TMGC researchers in screening ENU-mutagenized mice for phenotypes of interest and, ultimately, in using these mutant mice to help pinpoint the individual genes and gene products that are involved in the complex networks that create the phenotypic traits of interest. ***MuTrack ver 1.0*** has been completed and is routinely used in the TMGC (www.tnmouse.org/mutrack/index.php). The Web-based user interface allows researchers to enter, share, and analyze data about mice shipped around to several institutions and labs. A new system called ***Elector*** and ***GeneKeyDb*** is being designed to help find and analyze candidate genes in targeted chromosomal regions. In addition, another system, ***MuGnoSys ver 0.5*** allows researchers to develop complex and detailed knowledge about the different mutant mouse strains produced by TMGC or ORNL. These systems are being used to screen for neurological phenotypes but were developed to be scalable and adaptable for use in other rodent phenotype analyses, including detailed phenotypic analysis in mutants created by targeted mutagenesis or in the mutants found in the second approach discussed below.

ORNL is creating the ***Cyropreserved Mutant Mouse Bank (CMMB)*** of ENU-mutagenized mice that can be used to pursue a "gene-first" screening approach. This second approach is an alternative to the first "phenotype-first" screening approach. If bioinformatics and technology can help automate this screening, then large number of genes and SNPs can be screened at once and this can become an efficient method to create altered genes for further study. We have made progress in this bioinformatics automation. Additional comparative bioinformatics can also help prioritize testing of the discovered SNPs. If the screening is done on genomic sequence, for example, bioinformatics may suggest if a SNP could be in a possible transcription factor binding site (see analysis below) or if it is a SNP in a conserved protein-coding region that is likely to disrupt protein function.

Different alleles of a gene may perturb molecular aspects of these networked biosystems. Homeostatic regulation within these networks may tend to compensate for these perturbations and create different "molecular phenotypes' by affecting the expression of yet other genes in the network. RNA expression analysis (or proteome analysis) will help find these molecular phenotypes. Bioinformatics analysis of this molecular phenotype may help elucidate both additional participants in a network and the regulation of those networks. This analysis may also elucidate both the shared and different aspects of these networks in different cell types. ***GIMS ver. 1.0*** is a system under development for RNA expression data that can serve as a platform for further data mining and analysis developments. Combined work with ***MuTrack*** and ***GIMS*** is planned that would help track and analyze molecular phenotype data of interest. Complementary bioinformatics expertise and molecular phenotype data available in TMGC will be collaboratively applied to quantitative analysis of molecular phenotypic traits and variations. ORNL-based experimentalists are also studying the RNA expression of a set of genes in a keratinocyte model system.

Although different species exhibit a remarkable phenotypic diversity, recent genome analysis in bilaterally symmetrical animals suggests a remarkable similarity in the genes that are used in

important processes. Much biological diversity and complexity in the multicellular organisms may, in fact, be due to the evolution of gene regulatory networks. Comparative sequence analysis and studying the evolution of GRNs may help highlight both the conservation of underlying networks and their important differences. To begin this analysis, we are using several different sets of genes as test cases to develop an integrated and automated pipeline to analyze orthologs and paralogs in chordates. This may detect conserved genome sequences and sequences that are candidates for *cis*-regulatory sites for gene transcription. These *cis*-acting control regions are key integrators in gene regulatory networks. To narrow down regions for detailed study, we are developing tools to find and analyze large Conserved Genome sequence Blocks (CGSBs) at high throughput. Other analysis will be employed using these CGSBs to assist in better gene modeling and to suggest transcription factor binding sites. Several bioinformatics tools are being developed: these include a new phylogentic footprinting tool, ***GeneTreeConstructor, Lenny alignment tools***, and existing tools to be added to an integrated analysis pipeline. ***CGSBdb***, ***GeneKeydb,*** and ***GeneEvodb*** are data resources that are under design and construction. We are initially developing these resources to analyze the gene sets that are being explored in the keratinocyte RNA expression system in the hopes that we may be able to correlate the results from the computational and the experimental approaches.

Research sponsored in part by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory and the Office of Biological and Environmental Research and in part by a National Institutes of Health grant to the University of Tennessee. ORNL is managed by UT-Battelle LLC for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

94. The ORNL Genome Analysis Toolkit, Pipeline and DAS Server

Manesh Shah, Doug Hyatt, Frank Larimer, Philip LoCascio, Inna Vokler, and Edward C. Uberbacher
Oak Ridge National Laboratory
ube@ornl.gov

The Genome Analysis Toolkit (GAT) and Pipeline provide Internet configurable genome sequence analysis and annotation capabilities for both microbial and eukaryotic organisms. The GAT has been a major analysis engine for a large number of "microbe month" microbial genomes sequenced at the JGI, as well as for human, mouse, *Phanerochaete* (white rot fungus), and other microbes and eukaryotes. It provides a capability to present and update the JGI genome web pages and views of these many genomes in the ORNL Genome Channel. Usage of GAT outside ORNL has increased 20 fold in the last 12 months, such that the combined tools in the toolkit process upwards of 200 million bases per month (not counting ORNL), with major remote users including organizations such as the JGI, WUSTL, JAX, Accelrys, and Gene Logic.

We have continued to enhance and extend the capabilities of the Genome Analysis Toolkit with improvements to the individual tools incorporated in the toolkit, a redesigned the client-server system that provides access to the toolkit services, and significant refinements to the Web interfaces. In addition to interactive analysis, we have developed analysis pipelines to support comprehensive analysis of DNA sequences in "batch mode" for all supported genomes. These can be accessed at http://compbio.ornl.gov/tools/pipeline. A Java interface has been implemented for interactive, graphical visualization of the pipeline results. Data is also available via the newly implemented DAS (Distributed Annotation System) server at ORNL (http://genome.ornl.gov/das) which allows remote users to compare ORNL generated annotation with that developed at other sites.

The toolkit now includes a wide variety of genome analysis tools. Gene finding systems include new versions of GrailEXP (version 3.3) (http://compbio.ornl.gov/grailexp) for human, mouse, and *Phanerochaete*, Genscan for eukaryotic genomes, and Generation (http://compbio.ornl.gov/generation) and Glimmer for microbial gene prediction. Grail suite of tools includes CpG islands, PolyA sites and simple and complex repetitive elements, and BAC End identification. Also included are NCBI STS E-PCR, Pfam, RepeatMasker and tRNAscan-SE systems. Sequence similarity and protein domain search tools include

NCBI BLAST, Baylor Beauty post-processing, and the InterProScan analysis system.

The client-server system has been redesigned to facilitate handling of increased load on the system. The client-server protocol now incorporates issuance of a ticket (request ID) which eliminates the need for maintaining persistent client-server connections, allowing the client to retrieve the results later. The server distributes the incoming requests intelligently on the available pool of compute server machines, based on the loads on the various servers. Highly compute-intensive service requests (like BLAST, Pfam and RepeatMasker) are transparently redirected to the GIST (Genomic Integrated Supercomputing Toolkit) server running on ORNL's IBM RS/6000 SP supercomputer.

95. GrailEXP: Gene Recognition Using Neural Networks and Similarity Search

Doug Hyatt, Frank Larimer, Philip LoCascio, Victor Olman, Manesh Shah, Ying Xu, and Edward C. Uberbacher
Oak Ridge National Laboratory
ube@ornl.gov

GrailEXP 3.3 (October, 2001) represents the latest technology in genefinding. Many improvements have been made to GrailEXP over the past year, including the addition of alternative splicing recognition, the creation of systems for more eukaryotic organisms, the incorporation of GrailEXP into ORNL's high performance computing framework, and the provision of the executables to the academic/nonprofit community. Many future improvements are also under development, including a detailed protein homology search, a prototype genefinding system based on comparison between human and *Fugu rubripes*, an interactive Java interface for the GrailEXP suite, the addition of more organisms to the system, and the extension of the gene finding code to recognize more forms of alternatively spliced genes.

Each of the three codes in GrailEXP has been significantly enhanced since the last contractors' meeting. More Grail modules have been added to Perceval, the exon prediction code, including one for *Phanerochaete chrysosporium*, the white rot fungus genome sequenced by JGI. The accuracy of the alignments produced by Galahad, the EST alignment code, has dramatically improved; Galahad aligned known CDs entries in a test set against the genomic sequence with 99% accuracy. Finally, many significant additions have been made to Gawain, the gene modeling algorithm, including the recognition of alternatively spliced genes, "gluing" code to merge gene models that have been accidentally broken, and flexible rules sets for intron sizes/etc. for different eukaryotic organisms. Under development is a protein homology code based on BLASTX, a Java interface for the entire suite of tools, a more streamlined method for the automated training of new organisms, and a module based on aligning genomic sequences from different organisms (such as human, mouse, and *Fugu*.)

The public interface to GrailEXP has proven very successful. Many users are now regularly utilizing gene predictions from GrailEXP via one of three methods. The most popular still remains the Genome Channel (http://compbio.ornl.gov/channel/), which contains the results of running GrailEXP on the entire human and mouse genomes. In addition, the GrailEXP analysis page (http://compbio.ornl.gov/grailexp/) receives over 3000 requests per month from universities, nonprofit institutions, and companies around the world and processes and average of 63 million sequence bases per month, not counting requests for ORNL. Finally, the executables have been deployed at over 125 academic/nonprofit institutions. A publication for the latest version of GrailEXP is in preparation for 2002.

96. The Genome Channel: A Foundation for Genomes to Life and Comparative Genomics

Miriam Land, Frank Larimer, Jay Snoddy, Denise Schmoyer, Doug Hyatt, Manesh Shah, Inna Vokler, Philip LoCascio, Gwo-Liang Chen, Loren Hauser, and **Ed Uberbacher**
Computational Biology Program, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee
http://compbio.ornl.gov/
ube@ornl.gov

Representing DOE's unique perspective on eukaryotic and microbial genomes, the Genome Channel has made major improvements to its usability, organism coverage and supported types of analysis. Channel and related web sites receive an average of 25,000 web hits per day with usage comparing favorably with other major genome web resources. Genome Channel contains over 20 unique microbial genomes not contained on other sites and presents more types of human and mouse gene modeling than any other site. Integration of these many organisms in a single framework is making it possible to develop new tools and views that will help GTL researchers use comparative genomics to identify regulatory motifs and other conserved non-coding genome elements.

The Genome Channel has improved navigation tools which provide faster and easier access to the data. A new text-based home page has links for quick access to human and mouse data and an ever increasing list of both finished and draft microbials. Estimates of eukaryotic gene content in the Channel include five types: gene models based on GrailEXP, Genscan, Genbank annotation, U.Penn DOTS EST assemblies, and modeling based on Genbank Refseq mRNAs. Three of the methods include prediction of gene variants and these can be compared along with the evidence supporting each model. tRNA gene predictions have been added to both the text and graphical display. Microbial gene modeling is based on three methods; Generation, Glimmer, Critica, and an additional combined ORNL gene call based on these three methods. These suites of methods and organisms make the views provided for both eukaryotic and microbial genomes richer, more accurate, and more comprehensive than in other available resources. Many of the web pages now include links to 'lists' of data which can be downloaded and used with other tools. For example, lists of contigs, genes, or repeats on a chromosome or entire organism.

Annotation for JGI-sequenced draft microbials is accessible with links to several different summaries of the data which can be downloaded and put into other tools. These include Btab output for the predicted genes, Pfam summaries, and COGS analysis. Blast, Pfam, and COGs analysis can all be refreshed to get the latest interpretation of the data.

For GTL to be successful with Goal 2 (Gene Regulatory Networks), foundational data must be available that describes gene regulatory regions and potential transcription factor binding sites in genomes of interest. The most powerful paradigm for obtaining such information is through comparative genomics, since regulatory signals are often conserved among genomes, and also repeated within each genome proximal to genes of related function. Such inter- and intra-genome comparisons provide the basis for estimating regulatory signals and the role of genes in cell processes. The Channel and its underlying data warehousing structure and update processes are a natural foundation for facilitating comparison of genomes, including detailed comparison and analysis of regulatory units and structures in both microbes and eukaryotes. Steps are being taken to develop the Channel as a comprehensive resource for (1) facilitating and viewing large-scale genome alignment, (2) clustering and comparing gene/protein orthologs and paralogs, and identifying conserved blocks and potential regulatory motifs in their upstream regions, and (3) recognizing operon structure and comparing regulatory signals in operons of related species, and (4) identifying shared regulatory elements in regulons in genomes of interest to the GTL program.

(Research was supported by the Office of Biological and Environmental Research, USDOE, under contract number DE-AC05-00OR22725 with Oak Ridge National Laboratory, managed by UT-Battelle, LLC and through an ORNL Exploratory Studies Project.)

97. Automated Visualization of Large Scale Bacterial Transcriptional Regulatory Pathways

Carla Pinon, **Amit Puniyani**, Peter Karp, and Harley McAdams
Stanford University and SRI
amit8@stanford.edu

The objective of the Stanford Microbial Cell project is to identify the complete transcriptional regulatory network of the aquatic bacterium, *Caulobacter crescentus*. We are faced with the problem of comprehending the structure and function of these complex, highly interconnected networks. The EcoCyc system developed at SRI provides tools for

visualizing bacterial metabolic pathways in a visually appealing manner. We are building on the software underlying EcoCyc (called PathwayTools) to add capability to display and explore bacterial regulatory. By dynamically mapping gene expression levels determined by time sequences of RNA microarray assays onto the regulatory links, we can show the flow of control through the network as the cell cycle progresses or as the cell responds to changes in its environment. In this work we are taking advantage of methods and ideas previously used to illustrate social networks and the internet.

98. Integrating Computational and Human-Curated Annotations for the Mouse Genome

Carol J. Bult and the Mouse Genome Informatics Group
The Jackson Laboratory, Bar Harbor, ME, USA 04609
cjb@informatics.jax.org

Computationally-derived genome annotations are critical entry points into genome biology, however these also need to be integrated with existing biological knowledgebases to enable the research community to use the sequence information effectively in their research programs. The mission of the Mouse Genome Informatics (MGI) group (http://www.informatics.jax.org) is to evaluate computationally predicted gene models and integrate them with such information as genome location, alleles and phenotypes, homology, and gene expression patterns. The MGI group relies heavily on the expertise of domain specialists to analyze and interpret biological data from diverse sources as part of an overall strategy to create and maintain a highly integrated and well-curated genome biology database for the laboratory mouse[1]. A major challenge facing MGI, and all model organism databases, is how best to incorporate information from large and constantly changing genomic sequence data streams with curation processes that rely heavily on human reasoning and interpretation.

Within the MGI group we are meeting the large-scale genome annotation challenge by developing an annotation infrastructure that combines both automated and human-centric curation processes. I will present an overview of this infrastructure and discuss design and implementation issues relative to our current work on integrating and updating the annotations of the mouse full-length cDNA sequence data generated by the RIKEN group in Japan[2]. I will also present examples of our curatorial strategy for analyzing the available finished BAC sequences for the C57BL/6J strain of mouse that are available from GenBank and integrating these data with the wealth of genetic and biological knowledge about the laboratory mouse that is already available from MGI as well as other informatics resources.

1. Bult et al., 2000. *Proceedings of the IEEE International Symposium on Bio-Informatics and Biomedical Engineering:* 29-32.
2. The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium. 2001. *Nature* **409**: 685-690.

99. Comparative Sequence-Based Approach to High-Throughput Discovery of Functional Regulatory Elements

Gabriela G. Loots, Ivan Ovcharenko, Inna Dubchak, and Edward M. Rubin
LBNL/DOE/Genome Sciences, 1 Cyclotron Road, MS 84-171
ggloots@lbl.gov

A major challenge of the post-human sequencing era is identifying and decoding the regulatory networks underlining gene expression embedded in the large sea of noncoding DNA. While computational tools aimed at predicting regulatory elements identify large numbers of false positives, traditional experimental techniques, though more accurate, are slow and labor-intensive. We have developed a computational tool, *RegSeq: Regulatory Sequence Analyzer,* for high-throughput discovery of *cis*-regulatory elements that combines transcription factor binding site prediction (*TRANSFAC*) and the analysis of inter-species sequence conservation (global alignments). This process reduces the number of predicted transcription factor binding sites by several orders of magnitude, eliminating the majority of false positive hits. To illustrate the ability of *RegSeq* to identify true transcription factor binding sites we analyzed several AP-1, NFAT and GATA-3 experimentally characterized binding sites

in the 1 Mb well-annotated cytokine gene cluster (Hs5q31; Mm11). The exploitation of orthologous human-mouse data set resulted in the elimination of 95% of the 38,000 binding sites predicted upon analysis of the human sequence alone, while it identified 87% of the experimentally verified binding sites in this region. Since this region harbors a cluster of cytokine genes regulated by the GATA-3 transcription factor, we used *RegSeq* to analyzed the distribution of GATA-3 sites across the promoter regions of the 18 identified genes present on 1 Mb of orthologous human (5q31) and mouse (ch11) sequence. By searching the promoter sequences (2 kb upstream of the 5' UTR) for the presence of GATA-3 binding sites using *TRANSFAC* database we observed that the GATA-3 binding sites were abundantly and randomly distributed throughout the promoters of all 18 genes. We failed to observe a bias in the distribution of GATA-3 sites, favoring the genes known to be regulated by this transcription factor. Subsequent alignment of the human and mouse orthologous promoter regions revealed the presence of evolutionarily conserved GATA-3 sites only in the promoters controlling cytokine gene expression, the majority of which have previously been characterized as being GATA3 responsive. By combining sequence motif recognition provided by transcription factor database searches with multiple sequence alignment of orthologous regions, we have developed a high throughput strategy for filtering and prioritizing putative DNA binding sites involved in regulatory functions. The evolutionarily conserved transcription factor binding sites discovered by our study in the interleukin promoters serve as a genomically informed starting place for globally investigating the detailed regulation of this set of biomedically important genes.

100. Managing Targets and Reactions in a Finishing Database

Mark Mundt, Judith Cohn, Mira Dimitrijevic-Bussod, Marie-Claude Krawczyk, Roxanne Tapia, Al Williams, Larry Deaven, and Norman Doggett
DOE Joint Genome Institute and Center for Human Genome Studies, Los Alamos National Laboratory
mom@lanl.gov

The Los Alamos Center for Human Genome Studies is finishing human chromosome 16 employing a whole chromosome strategy. Using an Oracle database, single strand and low quality targets are maintained uniquely for the collection of overlapping projects on our minimal tiling path. While gaps remain in a project, a coordinateless definition of a target enables constant tracking of status to successful completion as well as eliminating the creation of duplicate reactions in different projects. One interesting case occurs when a new reaction is started within the boundaries of a target that must be split to record this history. Many SNP's and even multiple base pair differences are being documented in this manner of combining data in assemblies. The most important feature of this system; however, is that it provides an automated approach to supplying input/instruction lists for finishing robotics and operations employing the Q-Bot (cherry picking of subclones), Packard (cherry picking of DNAs), MerMades (96 channel oligonucleotide synthesizers) and 3700's (capillary sequencers) that keep our finishing efforts progressing. To date, we have tracked thousands of targets which have been successfully addressed.

Supported by the US DOE, OBER under contract W-7405-ENG-36.

101. Encoding Sequence Quality in BLAST Output by Color Coding

Sam Pitluck, Paul F. Predki, and Trevor L. Hawkins
U.S. DOE Joint Genome Institute, Walnut Creek, CA 94598
s_pitluck@lbl.gov

Tremendous amounts of draft sequence have been released into the public domain in recent years. While most of this sequence is of high quality, significant amounts of low quality sequence are still present. Because of this, draft sequence is of highest utility when the user is able to assess its quality. Although basecalling and assembly programs (such as Phred and Phrap) produce quality scores for each base, this information is typically lost by the time it reaches end users. We have added functionality to the BLAST program by color coding the output according to the quality scores. In another implementation, quality scores are directly displayed in the BLAST output. In either case, interpretation of BLASTing against draft sequence is significantly

simplified. Our public web servers now support color-coded BLAST.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, the Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098, and the Los Alamos National Laboratory under contract No. W-7405-ENG-36.

102. Whole Genome Assembly with JAZZ

Jarrod Chapman, Nicholas Putnam, and **Dan Rokhsar**
U.S. DOE Joint Genome Institute, Walnut Creek, CA, 94598
dsrokhsar@lbl.gov

We present a new graphical algorithm for whole genome assembly that self-consistently treats paired-end constraints. The algorithm was designed to be scalable for application to gigabase scale genomes, and has been parallelized using MPI. To aid in the development and validation of the assembler and its outputs, we have also developed a suite of graphical tools for examining and manipulating large assemblies. For typical 6X projects, large scaffolds are recovered, with sequence accuracy better than one part in 10,000. We describe the basic algorithm, demonstrate the visualization tools, and present results for a 30 MB fungal genome and a variety of microbes of varying sizes and depths sequenced at the JGI, and an initial assembly of the early whole genome shotgun sequencing of mouse.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, the Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098, and the Los Alamos National Laboratory under contract No. W-7405-ENG-36.

103. Assembly and Exploration of the Public Human Genome Working Draft

Terrence S. Furey[1], Jim Kent[2], and David Haussler[3]
[1]Computational Biology Laboratory, Department of Computer Science, University of California Santa Cruz
[2]Department of Biology, University of California Santa Cruz
[3]Howard Hughes Medical Institute, University of California Santa Cruz
booch@cse.ucsc.edu

A program written by UCSC student Jim Kent, called GigAssembler, is used to periodically assemble a widely used public draft version of the human genome sequence using updated data from GenBank at the National Center For Biotechnology Information (NCBI). This assembly is steadily improving as the public sequencing consortium churns out new data. We will look at the coverage statistics on the latest assembly, and then look at web tools to explore it, and what they find. The three most widely used public annotation browsers are the UCSC Genome browser (genome.ucsc.edu), the Ensembl genome browser (www.ensembl.org), and the NCBI map viewer (www.ncbi.nlm.nih.gov/genome/guide), the latter based on NCBI's own sequence assembly. We will focus on the UCSC browser, which shows a rich variety of data mapped to the genome sequence, including predicted genes, expressed sequence tags, full length mRNAs, genetic and radiation hybrid map markers, cytogenetically mapped clones, single nucleotide polymorphisms, homologies with mouse and pufferfish, and more. This data is presented on different tracks of annotation that are contributed by the annotation team at UCSC and more than a dozen researchers worldwide. We briefly discuss how web-based data browsers such as this are accelerating biomolecular and biomedical research, and how scientists and engineers in other disciplines can contribute to the study of the human genome.

104. Shotgun Sequence Assembly Algorithms for Distributed Memory Machines

Frank Olken
Lawrence Berkeley National Laboratory
olken@lbl.gov

This work is concerned with methods for computing shotgun sequence assemblies of distributed memory parallel computers, in which individual computing nodes can not contain the entire dataset.

For the overlap detection phase we use an algorithm modeled on distributive hash join algorithms, yielding a algorithm with linear speed up (with the number of processors).

For the layout phase we use a connected component labeling algorithm to partition the data for distinct connected components (i.e., the data for each connected component is contained in a single node). Each connected component can then be processed separately, in parallel using conventional layout algorithms. The overall work is linear in the problem size. Note that this method requires prior effective removal of repetitive DNA sequences.

105. Benefits of J2EE Architecture for Informatics Support of Genomic Sequencing

Roxanne Tapia, Judith Cohn, and Mark Mundt
DOE Joint Genome Institute and Center for Human Genome Studies, Los Alamos National Laboratory
rox@lanl.gov

At LANL, we have been building an informatics foundation for our next generation of Genomic Sequencing. Our primary goal is to provide integration in development, user interface, and data access for diverse components including, but not limited to, Laboratory Information Management Systems (LIMS), Quality Control, Sequence Analysis, Assembly and Annotation. We need an infrastructure that will allow quick adaptation to a dynamic, complex environment. Besides changeability, other characteristics of that environment include high-throughput, intensive processing, relatively few users on diverse operating systems, an existing Java codebase, lots of data, and a small bioinformatics team.

Given these characteristics, our challenge was to build an infrastructure that would allow us to support our current efforts and continue to evolve. Given a current Java codebase and skill set, combined with the openness of the Java Platform, and the versatility of the J2EE (Java 2 Enterprise edition) specification, Java was an easy choice for development. Because of the decision to leverage J2EE, we needed an application server that is J2EE compliant. We also needed a mature database that could support lots of data, and preferably one with Java and object support. Thus, Java, SilverStream Application Server, and Oracle RDBMS were chosen for our initial implementation.

Recently, we upgraded to the latest version of Oracle, and decided to replace SilverStream with Oracle 9i Application Server (9iAS). This transition has gone smoothly because we built our J2EE objects according to pure J2EE specifications, rather than using vendor-specific enhancements. This allowed us to switch application servers without an excessive migration cost. The poster will describe the benefits we have realized since adopting a J2EE architecture including flexibility of infrastructure, code reusability, and low cost of change.

Supported by the US DOE, OBER under contract W-7405-ENG-36.

106. Production Workflow Tracking and QC Analysis at the Joint Genome Institute

Heather Kimball, Stephan Trong, Art Kobayashi, Sam Pitluck, Yunian Lou, and Matt Nolan
U.S. DOE Joint Genome Institute, Walnut Creek, CA 94598
hlkimball@lbl.gov

The Joint Genome Institute Production Genomics Facility has produced over 2.75 billion bases of draft paired-end sequencing since January 1, 2001. Our sequencing methodologies incorporate two types of

DNA template generation: inoculation/SPRI purification and Rolling Circle Amplification. In order to manage the flow of samples through these processes, a robust database tracking system was developed using ORACLE. The key elements that are tracked within the workflow system include:

- Instruments
- Operators
- Protocols
- Reagents
- Dates and times
- Quality scores and contamination information

Data input and reporting for the workflow have been produced using a combination of commercial database development software and in-house programs. These include ORACLE's WebDB and Perl CGI programming. By leveraging the rapid report and form development cycle using WebDB and augmenting this with the flexibility of in-house programming, we have efficiently deployed a critical laboratory information management system for our data tracking.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, the Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098, and the Los Alamos National Laboratory under contract No. W-7405-ENG-36.

107. Goals, Design, and Implementation of a Versatile MicroArray Data Base

Marc Rejali[1], Marco Antoniotti[1], Vera Cherpinsky[1], Caroline Leventhal[1], Salvatore Paxia[1], Archisman Rudra[1], Joe West[2], and Bud Mishra[1]
[1]New York University Courant Bioinformatics Group
[2]Cold Spring Harbor Laboratory
mrejali@cat.nyu.edu

Many problems in functional genomics are being tackled using Microarray Technology. While this approach holds much promise for answering open questions in Biology, it poses significant problems from the "Data Management" point of view.

Our Bioinformatics group at NYU has been involved in several projects that use Microarray technology, for instance:

- Genome Mapping and Probe Placement (in cooperation with CSHL)
- Nitrogen Pathway analysis in Arabidopsis (in cooperation with NYU Biology Dept.)
- Hallucinogen effects on brain functions (in collaboration with Mount Sinai School of Medicine)
- Cancer related cell signaling using different cell lines (in cooperation with CSHL)

To address the needs of these collaborative research groups and others, we have developed the NYU Microarray Database (NYUMAD). It's functionality ranges from the storage of the data in relational data base management systems to front-end capabilities for the presentation and maintenance of the data.

The database is a unified platform to understand the microarray based gene expression data. The data can be output to a wide class of clustering algorithm, based on various "similarity measures" and various approaches to grouping. Particularly, we have developed a new statistically-robust similarity measure based on James-Stein Shrinkage estimators and provided a Bayesian explanation for its superior performance. Additional research is focused on incorporating statistical tests for validation and measuring the significance (e.g., jackknife and bootstrap tests). Finally, we plan to add an experiment design module, that suggests how the future array experiments should be organized, given that we understand how the past experiments have performed

Most of the underlying DB schema design follows closely the specifications put forth by the Microarray Gene Expression Database group (http://www.ebi.ac.uk/microarray/MGED), especially when it comes to the XML-based MAML exchange format.

Functionality: The functionality of the NYUMAD system is summarized hereafter:

- Microarray data is stored in relational data base management systems (RDBMS) using a database schema based on the MAML

(Microarray Mark-up Language) specification.

- Data is served to "clients" via the world wide web (WWW). Clients can be the NYUMAD Java applet that is part of the system described below, or custom-built user programs, or MAML XML files retrieved using a simple HTTP text based request format. In the case of the NYUMAD Java applet, data retrieval is generally transparent to the user and is carried out as a natural part of using the GUI front-end (see below). For text based requests, the returned data is in the MAML XML format.
- Data submissions for updating existing data or inserting new data can be made using the NYUMAD Java applet client, or by custom-built user programs, or HTML forms that access directly the server middle tier server . As with data retrieval, the GUI front-end capabilities of the NYUMAD Java applet make data submission transparent to the user.
- The NYUMAD applet presents data in a logical manner and allows easy navigation through the data. As the user navigates through the data, the required information is retrieved. It also allows straight-forward updating of existing data and the insertion of new data. The NYUMAD applet can also retrieve data in the MAML XML format which can then be cut and pasted to other applications.
- The NYUMAD system integrates several Clustering algorithms and libraries, in order to provide a complete service to the user. The integration is such to automatically access the Data Base and avoid tedious data reformatting and translation tasks.

Architecture: The NYUMAD has a three-tier architecture as shown in the diagram below.

- *Front tier* The Front tier comprises the NYUMAD applet and/or user's custom-built programs and HTML forms. The applet is written in Java and interacts with the middle tier using HTTP, requesting or submitting data using either a text based format or (Java) object serialization. Custom applications interact using HTTP and a text based format. The Microarray data in text based interactions is in MAML XML format.
- *Middle tier* The middle tier is provided by NYUMAD servlets written in Java that handle requests and submissions from the front tier. The middle tier is invisible to the end user. Requested data is retrieved from the RDBMS in the back tier using JDBC and then sent to the front tier either in MAML XML format or in the form of serialized objects for the NYUMAD applet or applications capable of interpreting the Java Object Serialization protocol. The middle tier servlets provide all the application logic necessary to ensure the integrity of the data and adherence to necessary rules, constraints and security restrictions. In addition the middle tier caches data, allowing for faster data retrieval and better scalability. The middle tier can access multiple back tier databases, allowing for data distribution and scalability. The middle tier uses the server's file management system to store and retrieve large files such as image files. In addition, the functional abstraction provided by the middle tier shields the front tier from changes in the back end structure, thus ensuring development extensibility and flexibility for the system.
- *Back tier* The back tier comprises the relational database management systems (RDBMS, currently PostgreSQL running on a 6 nodes Linux cluster) that store the Microarray and related data. It also includes the file management system used to store large files such as image files. The database schema is based on the MAML specification adapted to relational systems. Since the database access code in the middle tier uses JDBC, databases from different vendors can be used with relatively little additional code.

We have built the NYUMAD database as part of our integrated system for Bioinformatics centered on the VALIS tool. We took extreme care in making the system "distributable" from the start, by clearly defining a three tiered architecture that allows us to concentrate on different aspects of the design. We also closely followed the MAML standard put forth in the Spring of 2001 by the MGED group. It is our

intention to follow up on this design and to augment the NYUMAD DB with a module capable to communicate with other systems on the basis of the new MAGE-ML OMG Object model standard.

108. CLUSFAVOR – Computer Program for Cluster and Factor Analyses of Microarray-Based Gene Expression Profiles

L.E. Peterson
Departments of Medicine, Molecular and Human Genetics, and Scott Department of Urology, Baylor College of Medicine, Houston, Texas 77030

CLUSFAVOR is a Windows-based computer program for performing cluster and factor (principal components) analyses of microarray-based gene expression profiles. CLUSFAVOR was designed for the Windows 95/98/NT/2000/XP operating systems so that users can easily export images to Windows objects and/or JPG files for import into MS-Word and MS-Powerpoint. A significant amount of programming for CLUSFAVOR has focused on optimization and efficient use of resources to minimize run-times and memory usage. Recent enhancements to CLUSFAVOR in Version 5.0 include use of jackknife distance functions to reduce false positives due to outlier effects and speeding up eigenanalyses. This computer demo will include a tutorial on goals and assumptions of cluster and factor analyses, input data format, program usage, interpretation of results, and import of results into other Windows-based software. A CLUSFAVOR installation package and user guide can be downloaded from web page http://mbcr.bcm.tmc.edu/genepi.

Supported by NCI Supported by National Cancer Institute grant CA-78199-04.

109. Partitioning Large-Sample Microarray Transcription Profiles for Adaptive Response in Human Lymphoblasts Using Principal Components Analysis

L. E. Peterson[1], M. A. Coleman[2], E. Yin[2], B. J. Marsh[2], K. Sorensen[2], J. Tucker[2], and A. J. Wyrobek[2]
[1]Departments of Medicine, Molecular and Human Genetics, and Urology, Baylor College of Medicine, Houston, TX 77030
[2]Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94551
peterson@bcm.tmc.edu

We are currently using the Affymetrix U95A oligonucleotide array (12,626 genes) to detect differentially expressed genes for adaptive response in human lymphoblastoid cells. Major findings to date for cells given 2 Gy vs. cells given 2 Gy six hours after a priming dose of 5 cGy indicate ~2700 genes with marked changes across adapting and nonadapting conditions, 101 genes strongly upregulated, and 110 genes strongly downregulated under adapting conditions. This paper describes analytic research involving cluster and principal components analysis (PCA) to partition natural groupings of genes with similar transcription profiles for adaptive response. We used the CLUSFAVOR algorithm for cluster analysis and PCA. Because CLUSFAVOR uses the "foolproof" Jacobi method for eigenanalysis, significant program modifications were needed in order to reduce the prohibitively long calculation times normally required for Jacobi extraction of eigenvalues and eigenvectors from the large 12,626 x 12,626 (gene by gene) correlation matrix. Program enhancements resulted in substantial reduction of run-times from days to hours. More than 90% of total variance in input data was accounted for by extracting factors whose eigenvalues exceeded unity. Run-times for varimax orthogonal rotation were insignificant when compared with run-times for eigenanalysis. Bipolar factors containing strong positive and negative loadings were used to identify two unique groups of genes, since expression profiles of genes that load positive are the opposite of genes that load negative on the same factor. While cluster analysis generated a single dendogram (tree branch) containing every

gene in the input data, PCA assembled groups of genes with similar expression profiles. Image displays for cluster analysis and PCA of adaptive response transcription profiles will be presented. Statistical modeling of replicates and outliers will also be discussed.

[CLUSFAVOR algorithm development supported by National Cancer Institute grant CA-78199-04. Transcriptional profiling conducted under the auspices of NCI CA-DOE by the University of California, LLNL under contract W-7405-ENG-48 with support from NIH (ES09117-02) and DOE (KP110202).]

110. EXCAVATOR: Gene Expression Data Analysis Using Minimum Spanning Trees

Ying Xu, Dong Xu, Victor Olman, and Li Wang
Protein Informatics Group, Life Sciences Division, Oak Ridge National Laboratory
xyn@ornl.gov

Data clustering represents an essential step in large-scale data mining. While data clustering has been an active research subject for many years, a number of challenging problems remain, which have hindered in-depth applications of the technique. Among these challenges are (a) existing clustering methods generally do not guarantee globally optimal clustering results for multi-dimensional data; and (b) existing clustering methods often have problems with data sets containing clusters with complex boundaries. We have recently developed a new framework for multi-dimensional data clustering. The framework is based on a representation of multi-dimensional data as a minimum spanning tree (MST), a concept from the graph theory. A key property of this representation is that each cluster in the data set corresponds to one subtree of the MST representing the data, which rigorously converts a multi-dimensional clustering problem to a tree partitioning problem. We have rigorously demonstrated that though the inter-data relationship is greatly simplified in the MST representation, no essential information is lost for the purpose of clustering. Our research has led to the discovery of a number of highly attractive properties of MSTs for data clustering, which can help overcome some of the long-standing problems facing clustering techniques. Among them are that (1) the simple structure of a tree facilitates efficient implementations of rigorous clustering algorithms, which can guarantee the global optimality for clustering; and (2) the tree structure captures the essential topological information while leaving out the detailed geometric information among data points, which makes the cluster shape information irrelevant, and hence can handle clusters with extremely complex boundaries. Based on this framework, we have implemented a computer software EXCAVATOR for clustering and analyzing microarray gene expression data. By taking advantage of the key properties of MSTs, EXCAVATOR provides a number of unique features for gene expression data analysis, including (i) a capability of identifying significant clusters from a very noisy background; (ii) a capability of doing information-constrained clustering (e.g., genes X, Y, and Z should or should not belong to the same cluster); (iii) a capability of identifying genes with similar expression patterns to a set of seed genes; and (iv) a capability of automatic determination of the most plausible number of clusters in a data set. We have applied EXCAVATOR to a number of expression data sets from various genomes. The test results are highly encouraging.

References:

1. Y. Xu, V. Olman, and D. Xu, "Clustering Gene Expression Data Using a Graph-Theoretic Approach: An Application of Minimum Spanning Tree", *Bioinformatics*, 2001 (in press).
2. Y. Xu, V. Olman and D. Xu, "Minimum Spanning Trees for Gene Expression Data Clustering", *Proceedings of the Twelfth International Conference on Genome Informatics*, Dec. 2001 (in press).
3. D. Xu, V. Olman, L. Wang, Y. Xu, "EXCAVATOR: a software for gene expression data analysis", 2001 (submitted).

show that the EM algorithm performs efficiently and robustly.

- *Phase Detection by Optimization of an "Approximate Likelihood Function"*: to "contig" consecutive RFLP markers with a "phase" assigned to each RFLP in the contig. We develop a greedy strategy (based on local weighted averaging) that creates these contigs with relative phases assigned. Furthermore, we assign a "p-value" to each phased RFLP marker.

We have also developed a postscript-based visualizer that allows one to view the contigs globally and understand the structure of the haplotype blocks. We have experimented with a large amount of simulated data and the result so far has been gratifying and consistent with our theory. We plan to explore the performance of our algorithms (false-positive and false-negative RFLP markers and the size distributions of the haplotype blocks) as the optical mapping data deteriorates (e.g., decreased digestion rate, less-dense markers distributions, lower coverage by genomic DNA, etc.)

114. Information Management Infrastructure for the Systematic Annotation of Vertebrate Genomes

V. Babenko[1], B. Brunk[1], J. Crabtree[1], S. Diskin[1], Y. Kondrahkin[1], J. Mazzarelli[1], S. McWeeney[1], D. Pinney[1], A. Pizzaro[1], J. Schug[1], V. Bogdanova[2], A. Katohkin[2], V. Nadezhda[2], E. Semjonova[2], V. Trifonoff[2], N. Kolchanov[2], M. Bucan[3], and **C. Stoeckert**[1]
[1]Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA
[2]Institute for Cytology and Genetics, Novosibirsk, Russia
[3]Department of Psychiatry, University of Pennsylvania, Philadelphia, PA
stoeckrt@pcbi.upenn.edu

AllGenes (http://www.allgenes.org) is a human and mouse view of the GUS (Genomics Unified Schema) relational database and includes a gene index generated by assembly of publicly available EST and mRNA sequences. The assemblies integrate annotation from cDNA libraries, RH mapping data, and gene trap insertions. Automated annotation has been applied to characterize these sequences and relate them along with their predicted protein sequences to conceptual genes. The gene index contains over 3 million human and nearly 2 million mouse ESTs and mRNAs as of September, 2001 that have clustered into 150,006 human and 74,024 mouse "genes" (a new build of the index is underway). Approximately half the human and mouse genes have similarity to a known protein sequence and of these, we have been able to predict a Gene Ontology (GO) molecular function for 31% of the human and 45% of the mouse genes. Manual annotation is used to better structure the data (e.g., assign libraries to an anatomy ontology), confirm automated annotation (e.g., check GO assignments), and add new information (e.g., assign gene symbols and synonyms). Nearly 2000 human and mouse assemblies have been manually reviewed as of October, 2001 and this number is expected to greatly increase. Incorporation of genomic sequences will provide better assessment of the assemblies as genes and guide discovery of new genes and transcript alternative forms. The UCSC Golden path contigs are being used in this context with a focus on chromosome 22 of algorithmic and manual analysis. A related site on mouse chromosome 5 (http://www.cbil.upenn.edu/ mouse/chromosome5/) integrates the assemblies with existing genomic resources (e.g., BAC fingerprint, RH, and genetic maps) to facilitate functional analyses. The source and ownership of all data, algorithms run on it and evidence for assertions such as GO function predictions are stored in GUS allowing users to assess the validity of the data. The GUS schema is organized around the central dogma of biology (genes are transcribed to RNA which are translated to proteins) enabling a powerful query web interface. Users can build queries using Boolean functions to identify data sets for browsing and further analysis. The sequences, their contained accessions, predicted protein translations and predicted GO functions can be downloaded at the AllGenes site.

115. Manual Annotation of the Human and Mouse Gene Index: www.allgenes.org

Brian Brunk[1], Jonathan Crabtree[1], Sharon Diskin[1], **Joan Mazzarelli**[1], Chris Stoeckert[1], Nico Zigouras[1], Vera Bogdanova[2], Alexey Katohkin[2], Nikolay Kolchanov[2], Vorbjeva Nadezhda[2], Elena Semjonova[2], and Vladimir Trifonoff
[1]Computational Biology and Informatics Laboratory, Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104
[2]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
mazz@pcbi.upenn.edu

Allgenes.org is a web interface providing access to the assembled EST and mRNA sequences, or DoTS RNA transcripts, contained within GUS (Genomics Unified Schema), a relational database. The DoTS transcripts integrate annotation from cDNA libraries (tissue source) and RH mapping data also stored in GUS. Automated annotation has been applied to the DoTS transcripts to determine their predicted gene ownership, protein sequences and GO Functions. Manual annotation efforts have focused on validating the automated annotation and adding additional gene information. Manual annotation of the gene index utilizes an annotation tool, the GUS annotator interface, which directly updates the GUS database. Functional features of the interface which allow defined annotation tasks to be performed by the annotator include: determination of transcript gene membership using BLAST similarities and transcript alignments to genomic sequence, assignment of approved (HUGO or MGI) gene symbol, gene synonyms and confirmation/addition of protein GO Function assignments. Evidence for the automated annotation is stored in GUS and provided to the annotator to assist in the validation of the assignments. Evidence is also manually added by the annotator for each assignment and is stored in GUS. The human DoTS transcripts have been aligned on the UCSC Golden path contigs allowing for the identification of new genes, alternative transcript forms and annotation of the genome. Manual annotation efforts have focused initially on the genes contained within the region deleted on chromosome 22, causing DiGeorge syndrome, a developmental disorder.

116. The Comprehensive Microbial Resource

Owen White, Lowell Umayam, Tanja Dickinson, and Jeremy Peterson
The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850
owhite@tigr.org

One of the challenges presented by large-scale genome sequencing efforts is the effective display of information in a format that is accessible to the laboratory scientist. The Comprehensive Microbial Resource (CMR) contains all of the fully sequenced microbial genomes, the curation from the original sequencing centers, and further curation from TIGR (for those genomes sequenced outside TIGR). The interface to this database effectively "slices" the vast amounts of data in the sequencing databases in a wide variety of ways, allowing the user to formulate queries that search for specific genes as well as to investigate broader topics, such as genes that might serve as vaccine and drug targets. The web presentation of the CMR includes the comprehensive collection of bacterial genome sequences, curated information, and related informatics methodologies. The scientist can view genes within a genome and can also link across to related genes in other genomes. The effect is to be able to construct queries that include sequence searches, isoelectric point, GC-content, GC-skew, functional role assignments, growth conditions, environment and other questions, and isolate the genes of interest. The database contains extensive curated data as well as pre-run homology searches to facilitate data mining. The interface allows the display of the results in numerous formats that will help the user ask more accurate questions. The methodology for populating the database, the user interface, and new methods for automated analysis will be presented.

Microbial Cell Project

117. The Molecular Basis for Metabolic and Energetic Diversity

Timothy Donohue, Jeremy Edwards, Mark Gomelsky, Jon Hosler, Samuel Kaplan, and William Margolin
University of Wisconsin-Madison Department of Bacteriology, University of Delaware Department of Chemical Engineering, University of Wyoming Department of Molecular Biology, University of Mississippi Department of Biochemistry, and University of Texas Medical Center at Houston Department of Microbiology and Molecular Genetics
tdonohue@bact.wisc.edu

Our long-term goal is to understand and to capitalize on the metabolic activities of facultative microorganisms. To accomplish this, we will be acquiring a comprehensive understanding of metabolic pathways, bioenergetic processes, and genetic regulatory networks in the facultative phototroph *Rhodobacter sphaeroides*, strain 2.4.1. No single organism has the number of variety of metabolic abilities that are known to exist in this α-proteobacterium. The *R. sphaeroides* genome sequence predicts the existence of additional DOE-relevant pathways that were not thought to exist in this organism (http://genome.ornl.gov/microbial/rsph/). The known metabolic potential of *R. sphaeroides* includes,

- photoheterotrophic growth with light in the absence of O_2 using a variety of organic carbon compounds as a source of carbon and reducing power,
- photoautotrophic growth with light in the absence of O_2 using CO_2 as sole carbon source and H_2 as a source of reducing power,
- chemoheterotrophic growth without light in the presence of O_2 using a variety of reduced organic compounds as a source of carbon and reducing power,
- chemoautotrophic growth without light in the presence of O_2 under using CO_2 as sole carbon source and H_2 as a source of reducing power, and
- fermentative growth without light in the absence of O_2.
- the ability to degrade organic or inorganic toxins (including heavy metal oxyanions and oxides).

In this project, the Microbial Cell Project (http://microbialcellproject.org/) is supporting a team of microbiologists, biochemists, and metabolic engineers to analyze and model the flux of metabolites, the components of bioenergetic pathways, the assembly of key bioenergetic complexes, and the linkages between energetic and global regulatory networks in this facultative bacterium. This poster will describe the major projects that have been initiated since this project began in September, 2001. It will provide an overview of efforts we have begun to

- analyze the 5 predicted terminal oxidases of *R. sphaeroides*. This work includes efforts to characterize the biochemical and bioenergetic features of each enzyme, to develop models that predict the flow of reducing power through each branch of the aerobic respiratory chain at different O_2 tensions, and to use a defined set of oxidase mutants to determine the contribution of each oxidase to aerobic energy generation.
- define the role of newly discovered components of the photosynthetic electron transport chain. This work includes the use of mutant strains and metabolic engineering principles to determine how reducing power from the cyt bc_1 complex is funneled to the photochemical reaction center, the cyt cbb_3 oxidase, or other electron acceptors.
- dissect the genetic regulatory networks that control the expression of key suites of genes. This work includes identifying the role of O_2-sensitive transcription factors like FnrL, signal transduction pathways such as PrrABC, or alternative sigma factors in regulating expression of electron carriers that are needed to generate energy, assimilate CO_2/N_2, or remove toxic

compounds. Conditions are also being optimized for using DNA microarray technology to identify genes that are controlled by potential hierarchies of global regulatory networks.

- analyze the levels and turnover of metabolites that contribute to cellular energy generation in both steady state cultures and one that are adapting to new metabolic conditions. This work includes analysing those metabolites that set the energetic state or oxidation-reduction potential of the cell.
- analyze assembly of critical bioenergetic enzymes. This work includes using immunofluorescence microscopy to visualize the formation of bioenergetic machines like the photosynthetic apparatus in both wild type cells and mutants that lack presumed assembly factors.

118. Genome Sequence-Based Functional and Structural Analysis of a Transformable Cyanobacterium: the *Synechocystis sp. PCC 6803* Microbial Cell Project

Wim Vermaas[1], Robert Roberson[1], Martin Hohmann-Marriott[1], Daniel Jenk[1], Zhi Cai[1], Kym Faull[2], and Julian Whitelegge[2]
[1]Department of Plant Biology and the Photosynthesis Center, Arizona State University, Box 871601, Tempe, AZ 85287-1601
[2]Pasarow Mass Spectrometry Laboratory, Departments of Chemistry and Biochemistry, Psychiatry and Biobehavioral Sciences and The Neuropsychiatric Institute, UCLA, 405 Hilgard Avenue, Los Angeles, California 90095
wim@asu.edu

A completed genome sequence is only an excellent start: the next challenge is to determine the functional relevance of open reading frames for the physiology of the organism. To enhance the potential for functional analysis of open reading frames, a convenient transformation system should be available, and mutants impaired in key metabolic processes should be able to survive under specific conditions. These properties, and more, are found in the cyanobacterium *Synechocystis* sp. PCC 6803. This organism was the first phototroph of which a genomic sequence was determined, and it is spontaneously transformable, integrates DNA by double-homologous recombination, and can grow in the absence of photosynthesis or respiration.

Cyanobacteria are close to the ancestors of chloroplasts, but cyanobacteria maintain fully functional respiratory and photosynthetic complexes in the same internal membrane system, the thylakoids. Several redox-active components and proteins are shared between photosynthesis and respiration. A large number of *Synechocystis* mutants has been generated that are altered in one of the photosystems or in one or more of the respiratory complexes, and the properties of the various strains are being compared using a range of mostly *in vivo* techniques. These techniques include (1) redox measurements of the plastoquinone pool, (2) electron transport rate measurements using an oxygen electrode, (3) chlorophyll fluorescence analysis indicating the redox state of the system, (4) localization of protein complexes by means of fluorescent tagging, (5) ultrastructural analysis of membranes and cells, eventually leading to a 3-dimensional view of the cell, (6) quantitative analysis of metabolites such as organic acids using GC/MS, and (7) comparative proteomics of different mutants, the subject of the presentation by Whitelegge et al.

The *Synechocystis* Cell Project focusing initially on photosynthetic and respiratory processes has been started in September 2001, and builds on very significant insight that has been obtained since the genome sequence was completed. This insight includes, among others: (1) Cells can be grown in the absence of either oxygen evolution (by photosynthesis) or oxygen uptake (by respiration) when a fixed-carbon source is available; the identity of the electron acceptor in this case is being investigated. (2) Even though a 2-oxoglutarate dehydrogenase complex is missing according to the genome sequence, a complete TCA cycle has been demonstrated; an alternate shunt between 2-oxoglutarate and succinate is proposed. (3) A traditional membrane subunit of succinate dehydrogenase is missing according to the genome sequence; however, a very different subunit resembling that from Archaea is present and appears functional, suggesting a lateral gene transfer event. (4) The plastoquinone pool, a central redox buffer in both photosynthesis and respiration, is fairly reduced

in darkness and is oxidized in the light; this is consistent with a much higher abundance of photosystem I (taking electrons from the pool) than photosystem II (donating electrons to the pool) in this cyanobacterium. (5) Pigment cofactors, including chlorophylls and carotenoids, can be altered somewhat in their chemical structure without major functional consequences; this illustrates the plasticity of pigment binding sites and of pigment function.

Together, the results obtained thus far on *Synechocystis* sp. PCC 6803 illustrate the major impact a genomic sequence can have on starting to understand the molecular physiology of a cell, particularly when the genomic sequence has been obtained on an organism that is easily accessible to targeted genetic modifications.

The *Synechocystis* Microbial Cell Project is funded by DOE (DE-FG03-01ER15251).

119. A Pathway/Genome Database for *Caulobacter crescentus*

Pedro Romero[1], William Lee[2], Alison Hottes[2], and **Peter D. Karp**[1]
[1]Bioinformatics Research Group, SRI International, 333 Ravenswood Ave, EK207, Menlo Park, CA 94025
[2]Stanford University
pkarp@ai.sri.com

We have used SRI's Pathway Tools software to predict the metabolic-pathway complement of *Caulobacter crescentus*. The resulting prediction is stored within a pathway/genome database that integrates information about the genes, gene products, and biochemical pathways of *Caulobacter*. The poster will describe the computational method for predicting metabolic pathways, and the Pathway Tools components for querying and visualizing pathway/genome databases. The poster will also summarize what pathways were identified in *Caulobacter*.

120. Characterization of Genetic Regulatory Circuitry Controlling Adaptive Regulatory Pathways in a Bacterial Cell

Harley McAdams[1], Michael Laub[2], Peter Karp[3], Lucy Shapiro[1], Alfred Spormann[1], and Charles Yanofsky[1]
[1]Stanford University
[2]Harvard University
[3]SRI International
mcadams@cmgm.stanford.edu

An interdisciplinary team from Stanford, Harvard, and SRI International is identifying and characterizing the complete genetic regulatory circuitry and metabolic pathways of the aquatic bacterium, *Caulobacter crescentus*. The global physiological responses of *C. crescentus* cells and cultures during starvation, during adaptation to exposure to toxic chemicals, during exposure to alternative, environmentally-relevant catabolic substrates, and in biofilms will be determined using DNA microarrays, metabolic biochemistry and metabolic profiling, pathway and circuitry modeling, and bioinformatics.

The initial step is to estimate the overall regulatory and metabolic networks in *C. crescentus* through gene expression microarray assays of the wild type strain and bioinformatics analysis. For example, gene expression of wild type cells will be assayed for a selected set of time courses when subjected to diverse environmental conditions, such as different nutrient levels, transition into and out of stationary state, sudden exposure to several stresses and to changed nutrients in the environment, and growth as biofilms. Cluster analysis and analysis of the timing patterns in these datasets will predict sets of genes that are regulated as cascades or cassettes. From sequence homologies and the temporal patterns, candidate regulatory genes will be identified. Then the Pathway Tools software will be used to identify *C. crescentus* metabolic pathways. Additional microarray studies as well as conventional genetic and biochemical analyses of both wild type and mutant strains will then be designed to verify the postulated regulatory network.

Initial results have identified the operon organization of the *C. crescentus* genome and time varying gene expression levels from synchronized cultures of both wild type and mutant cells. Also, tools to visualize and "browse" the genome structure have been constructed.

121. Transcription Unit Organization and GAnTC Site Distribution — Two Studies of Genome Organization in *Caulobacter crescentus*

Alison Hottes, Swaine Chen, Lucy Shapiro, and Harley McAdams
Stanford University
ahottes@stanford.edu

With the increasing availability of both fully sequenced microbial genomes and microarray data, it is becoming easier to consider the overall organization of bacterial genomes. As part of the Stanford Microbial Cell Project, we have looked at two aspects of the aquatic bacterium *Caulobacter crescentus*' genomic organization – its transcription unit (operon) structure and the distribution of GAnTC methylation sites within its genome.

1. The operon-level organization of a bacterial genome forms the foundation of that bacterial cell's transcriptional regulatory network. We have used a probabilistic framework to predict the operon organization of the *C. crescentus* genome. Our computational method uses microarray data from both mutant and cell cycle studies, information about adjacent pairs of genes conserved across multiple bacterial genomes, and genomic spacing data. The algorithm makes extensive use of information about the size and spacing of transcription units in the more widely studied *Escherichia coli*.
2. In *Caulobacter crescentus*, CcrM is an essential cell cycle-regulated DNA methyltransferase that methylates the adenine in GAnTC sites. CcrM is not part of any known restriction/modification system. In the *C. crescentus* genome, GAnTC sites appear less frequently than would be expected by chance alone and are distributed non-uniformly around the chromosome. We present a variety of statistical studies, including correspondence analysis, that were conducted to determine a function for CcrM.

This research is part of the Stanford University Microbial Cell project. The objective of the project is to identify the complete transcriptional regulatory network of *Caulobacter crescentus*.

References

1. Craven M, Page D, Shavlik J, Bockhorst J, Glasner J. A probabilistic learning approach to whole-genome operon prediction. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2000; 8: 116-27.
2. Ermolaeva MD, White O, Salzberg SL. Prediction of operons in microbial genomes. *Nucleic Acids Res.* 2001 Mar 1; 29(5): 1216-21.
3. Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Paley, S. and Pellegrini-Toole, A. EcoCyc: Electronic Encyclopedia of *E. coli* Genes and Metabolism. *Nucleic Acids Research*, 28(1): 56 2000.
4. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J. Operons in *Escherichia coli*: genomic analyses and predictions. *Proc Natl Acad Sci USA.* 2000 Jun 6; 97(12): 6652-7.
5. Stephens C, Reisenauer A, Wright R, Shapiro L. A cell cycle-regulated bacterial DNA methyltransferase is essential for viability. *Proc Natl. Acad. Sci. USA.* 1996 Feb 6; 93(3): 1210-4.

122. Genome-Wide Survey of Protein-Protein Interactions in *Caulobacter crescentus*

Peter Agron and Gary Andersen
Biology and Biotechnology Research Program,
Lawrence Livermore National Laboratory,
Livermore, CA 94551
agron1@llnl.gov

We have recently initiated a study aimed at taking advantage of the complete genome sequence of

Caulobacter crescentus to study protein-protein interactions as a means to help elucidate genome function. As well as serving as a model system to study the cell cycle and cellular differentiation, this experimentally tractable gram-negative bacterium flourishes in aqueous environments with dilute nutrients, making it an attractive bioremediation vehicle. Our approach uses a commercially available two-hybrid system (Stratagene) in *Escherichia coli.* Vectors that encode fusions to two protein fragments (bacteriophage λ cI and the N-terminus of the RNA polymerase α-subunit) confer a biological activity that can be easily assayed (penicillin resistance) if the two chimeric proteins bind. We plan to adapt this system to allow up to approximately 200 genes of interest to be tested for interactions using a random-fragment library. Here we present pilot studies to test the efficacy of the approach using genes known to encode interacting proteins. Interactions were tested for structural proteins involved in cell division (FtsZ/FtsZ, FtsZ/FtsA, Smc/Smc), regulatory proteins involved in chemotaxis (CheA/CheW and CheA/McpA), regulatory proteins involved in the cell cycle and differentiation (CtrA/DivK), and regulatory proteins involved in nutrient uptake (PhoR/PhoB). Several of these genes encode two-component histidine kinases and response regulators, an important class of signal transduction proteins that is very abundant in *C. crescentus* and crucial for many cellular processes. The results of the pilot studies and the outlook for large-scale experiments will be discussed.

123. Relationship Between Metabolism, Oxidative Stress and Radiation Resistance in the Family *Deinococcaceae*

Amudhan Venkateswaran[1], **Marina Omelchenko**[2], Hassan Brim[1], and Michael J. Daly[1]
[1]Uniformed Services University of the Health Sciences (USUHS) 4301 Jones Bridge Road Department of Pathology, Room B3153, Bethesda, MD 20814-4799
[2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894
Avenkateswaran@usuhs.mil; omelchen@ncbi.nlm.nih.gov

Bacteria belonging to the family *Deinococcaceae* are some of the most radiation resistant organisms discovered. Different species of the *Deinococcus* genus were examined for their metabolic diversity and resistance to oxidative stress and radiation. Of the known deinococcal species, the thermophilic *Deinococcus geothermalis* is the least resistant and the only one endowed with complete biosynthetic capabilities. By comparison, the other members of the genus show greater resistance, but are highly dependent on cysteine and nicotinic acid for growth, and secrete proteases. Following the recent sequencing and annotation of *D. radiodurans*, research on its resistance capabilities is expanding to include the consideration of global cellular processes within which protection and repair systems operate efficiently. We present data consistent with the hypothesis that metabolism of *D. radiodurans* has evolved to minimize production of oxygen free radicals. While defects in deinococcal metabolism severely limit growth in nutritionally restricted environments, they likely enhance recovery of cells with accumulated DNA damage.

124. Structural Analysis of Proteins Involved in the Response of *Deinococcus radiodurans* to DNA Damage

Stephen Holbrook[1], Ursula Shulze-Gahmen[1], David Wemmer[1], James Berger[1], Sung-Hou Kim[1], Steven Brenner[1], and Michael Kennedy[2]
[1]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
[2]Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA 99352
srholbrook@lbl.gov

Deinococcus radiodurans (*Dr*) has the most efficient DNA repair and maintenance system of any organism yet identified. Insights into the general nature of DNA repair in all organisms may be achieved through detailed structural analysis of the proteins involved in the repair pathways of this organism. Current research is focused on two important components of the *Dr* DNA-repair system, the recFOR pathway that facilitates repair of double-

strand DNA breaks by homologous recombination, and the Nudix family of nucleotide polyphosphate hydrolases, for which 21 distinct proteins are present in *Dr*, the most in any organism. Some members of the Nudix family limit mutations by hydrolyzing oxidized products of nucleotide metabolism that are mutagenic when incorporated into the genome.

The genes constituting the recFOR pathway have been successfully cloned in our lab from *Dr* as well as from the hyperthermophilic organism, *Thermus thermophilus*. Recombinant RecO and RecR are well-expressed soluble proteins while RecF tends to from inclusion bodies in bacterial overexpression systems. Purification and crystallization experiments of the soluble RecO/R proteins are under way and different expression systems for RecF will be tested in the future.

Bacterial clones for six recombinant Nudix proteins from *Dr* were obtained from M. Bessman (Johns Hopkins) and we have cloned three other Nudix proteins. All of the proteins are soluble and express well and they are currently being purified. One of the Nudix proteins (DR1025) was crystallized and X-ray diffraction data collected to high resolution (1.6Å). Crystallographic phasing and structure solution is under way. Another Nudix protein (DR0079) is under analysis by NMR methods.

125. The *Deinococcus radiodurans* Microarray: Changes in Gene Expression Following Exposure to Ionizing Radiation

John R. Battista[1], Heather A. Howell[2], Mie-Jung Park[1], Ashlee M. Earl[1], and Scott N. Peterson[2]
[1]Louisiana State University and A & M College, Baton Rouge, LA 70803 USA
[2]The Institute for Genomic Research, Rockville, MD 20850 USA
jbattis@lsu.edu

The *D. radiodurans* genome encodes essentially the entire ensemble of DNA repair proteins found in *E. coli*. With the exception of alkylation transfer and photoreactivation, all of the major prokaryotic DNA repair pathways are represented. This observation is significant, not because it says anything about why *D. radiodurans* is radioresistant, but because it confirms something long suspected; *D. radiodurans* possesses unique mechanisms for dealing with ionizing radiation-induced DNA damage. Clearly, the collection of identified repair proteins in *D. radiodurans*, in and of itself, is not sufficient to confer radioresistance. If it were, *E. coli* would be as radioresistant. Assuming that the identified repair proteins encoded by *D. radiodurans* perform the same functions as their *E. coli* homologues, it seems reasonable to expect that *D. radiodurans* expresses novel proteins that enhance this species survival. Of the 3187 open reading frames identified in *D. radiodurans* R1, only 1493 could be assigned a function based on similarity to other gene products found in the protein databases. Of the 1694 proteins of unknown function, 1002 are, at present, unique to *D. radiodurans*, showing no database match to any other previously sequenced gene. The secret to understanding the radioresistance of *D. radiodurans* is presumably found among these proteins of unknown function. To achieve the goal of defining the proteins necessary for the radioresistance of *D. radiodurans*, we have constructed a DNA microarray, and used this array to follow changes in gene expression as this species recovers from a sub-lethal dose (3000Gy) of ionizing radiation. Fewer than 50 loci showed significant increases in expression in response to this treatment during the first hour of recovery and approximately half of the induced genes encode proteins of unknown function. In a parallel study, R1 cultures were desiccated for a period of two weeks and gene expression followed during rehydration. The pattern of gene expression following desiccation was compared with that observed following ionizing radiation because desiccation, like ionizing radiation, induces DNA double strand breaks. The patterns of gene expression partially overlapped, identifying five loci (DR0003, DR0070, DR0326, DR0423, DRA0346) encoding hypothetical proteins. This result suggests that these loci are involved in *D. radiodurans*' tolerance of DNA double strand breaks.

126. A Conceptual and In Silico Model of the Dissimilatory Metal-Reducing Microorganism, *Geobacter sulfurreducens*

Derek R. Lovley, Madellina Coppi, Stacy Cuifo, Susan Childers, Ching Lean, Franz Kaufmann, Daneil Bond, Teena Mehta, and Mary Rothermich
Department of Microbiology, University of Massachusetts, Amherst, MA 01003
dlovley@microbio.umass.edu

The long-term objective of this project, which has just begun, is to develop comprehensive conceptual and mathematical models of *Geobacter* physiology, and the interaction of *Geobacter* with its physical-chemical environment, in order to predictively model the behavior of *Geobacter* in a diversity of subsurface environments. Initial studies are focusing on Geobacter sulfurreducens because: 1) this microorganism has all of the unique physiological properties characteristic of *Geobacter* species; 2) the complete genome sequence is available; 3) a genetic system is available; 4) DNA-microarrays for the complete genome will be available; 5) it can be grown in chemostats; and 6) complementary functional genomic studies are underway.

In the first three years our objectives are to understand and model: 1) central metabolism in *G. sulfurreducens*; 2) electron transport to Fe(III) oxide, the electron acceptor supporting the growth of *Geobacter* in the subsurface; 3) electron transport to U(VI); 4) growth under the energy- and nutrient-limited conditions found in subsurface environments; 5) general regulatory mechanisms; and 6) response to environmental stresses such as oxygen and toxic metals. Chemostat studies are underway to collect physiological data to be compared with the predictions of the in silico model. Preliminary results on acetate uptake in the chemostats indicate that *G. sulfurreducens* can readily metabolize acetate at the low concentrations (ca. 10 μM) found in subsurface environments. Growth yields with fumarate as the electron acceptor are twice those with Fe(III) as the electron acceptor. This indicates that there may be an additional proton-translocation step in electron transport to fumarate. This hypothesis is being further evaluated.

Analysis of the genome predicted a number of previously unknown physiological characteristics which were evaluated in vivo. For example, putative chemotaxis genes were identified and subsequent investigations demonstrated that *G. metallireducens* was chemotactic toward Mn(II) and Fe(II), which are likely to guide the organism to Mn(IV) and Fe(III) oxides under anaerobic conditions. This is the first example of chemotaxis to metals. Although previous studies suggested that *G. sulfurreducens* and other *Geobacter* species are strict anaerobes, the genome contained genes coding not only for proteins involved in oxygen tolerance, such as catalase, superoxide dismutase, rubrerythrin, three potential rubredoxin genes, and a homolog of NADH/rubredoxin oxidoreductase, but also a homolog of a gene involved in oxygen reduction in *Desulfovibrio spp.*, rubredoxin oxygen oxidoreductase.

Further physiological analysis demonstrated that *G. sulfurreducens* could recover from oxygen exposure for more than a day and we are currently investigating whether it is able to use oxygen as an electron acceptor in a manner similar to another "strict anaerobe", *Desulfovibrio*. One of the more unexpected and fascinating stories of the *G. sulfurreducens* genome is the finding that *G. sulfureducens*' metabolism is likely to be much more highly regulated than was hypothesized prior to sequencing the genome. Of particular interest are putative iron-sensing systems that are likely to be involved in controlling expression of key metabolic genes. The role of these and other putative regulator proteins are being intensively investigated.

In the future, the developing in silico model will be used to aid in the design of experimental approaches and as more information on gene function and regulation becomes available this will be incorporated into the developing in silico model. It is hypothesized that through this iterative approach it will be possible to evolve an in silico model that can predict the growth and activity of *G. sulfurreducens* in chemostats and, eventually, in the subsurface. In addition to making numerous contributions to the basic understanding of microbial physiology, a model capable of predicting the growth and activity of *Geobacter* species in the subsurface will greatly aid in the design of rational strategies for the in situ bioremediation of metal and organic contamination.

127. The *Rhodopseudomonas palustris* Microbial Cell Project

F. Robert Tabita[1], Janet L. Gibson[1], J. Thomas Beatty[2], James C. Liao[3], Caroline S. Harwood[4], Timothy D. Veenstra[5], Frank Larimer[6], Joe (Jizhong) Zhou[6], and Dorothea Thompson[6]
[1]The Ohio State University
[2]University of British Columbia
[3]Univeresity of California at Los Angeles
[4]University of Iowa
[5]Pacific Northwest National Laboratory
[6]Oak Ridge National Laboratory
tabita.1@osu.edu

The objective of the *Rhodopseudomonas palustris* Microbial Cell Project is to examine how processes of global carbon sequestration (CO_2 fixation), energy generation from light, biofuel (H_2) production, plus organic carbon catabolism and aromatic hydrocarbon degradation and metal reduction operate in a single microbial cell. The recently sequenced *Rhodopseudomonas palustris* genome serves as the raw material for these studies. The control of all these processes appears to be integrated and the major aim of this study is to determine how this integrative metabolism is controlled and how certain aspects of metabolism, such as energy (H_2) production and carbon sequestration, might be enhanced in this versatile organism. We have assembled a team of investigators, from four academic and two DOE national laboratories, who share a common interest in bringing diverse approaches and types of expertise to bear on this important problem. Coordinated application of gene expression profiling, proteomics, carbon flux analysis and bioinformatics approaches are combined with traditional studies of mutants and physiological/biochemical characterization of cells. More specifically, functional analysis of the *R. palustris* proteome and global gene expression is considered within the context of the biochemistry and physiology of interactive control of energy generation and aerobic/anaerobic CO_2 assimilation and N_2 fixation, H_2 oxidation and H_2 evolution, and sulfur metabolism. These studies take advantage of the fact that *R. palustris* is phototrophic, can fix nitrogen, and may evolve copious quantities of hydrogen gas; the organism is unique in its ability to use a diversity of substrates for both autotrophic CO_2 fixation (i.e., H_2, H_2S, $S_2O_3^{2-}$, formate) and heterotrophic carbon metabolism (i.e., sugars, dicarboxylic acids, and aromatics, plus many others) under both aerobic and anaerobic conditions. As the project develops, intracellular localization and modeling of the expression of the key cellular processes will be undertaken.

128. Development of a DNA Microarray to Characterize the Roles of Apparently Redundant Genes in *Rhodopseudomonas palustris*, a Versatile Phototroph

Caroline S. Harwood[1], Dorothea Thompson[2], and Jizhong Zhou[2]
[1]Department of Microbiology, University of Iowa, Iowa City, IA
[2]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN
caroline-harwood@uiowa.edu

Rhodopseudomonas palustris is a very successful photosynthetic bacterium that can be found in virtually any temperate soil or water sample on earth. It is among the most metabolically versatile of known bacteria and has many alternative ways of acquiring carbon and nitrogen and of generating energy. It can grow anaerobically in light, convert many diverse kinds of aromatic compounds to cell material, and convert gaseous nitrogen to ammonia. Each of these aspects of the biology of *R. palustris* is reflected in its 5.49 Mb genome and its 5,000 potential protein encoding genes. Moreover, *R. palustris* has apparently functionally redundant genes to encode each of these processes. As part of a project to use whole genome microarrays to define the genes and molecular regulatory mechanisms that are responsible for the metabolic versatility of *R. palustris*, we have begun construction of a pilot microarray. The microarray will be used to identify physiological conditions under which apparently functionally redundant genes involved in light harvesting, nitrogen fixation and aromatic compound/fatty acid degradation are differentially expressed relative to each other. The pilot microarray will also be used to identify regulatory genes that control the differential expression of each of these processes.

129. Global Characterization of Proteins Associated With *S. oneidensis* MR-1 Outer Membrane Vesicles

Margaret F. Romine[1], Jim Fredrickson[1], Yuri Gorby[1], Jeff McLean[1], Mary S. Lipton[1], Ljiljana Pasa-Tolic[1], Alexander Tsapin[2], Kenneth Nealson[3], Carol Giometti[4], Sandra Tollaksen[4], and Richard D. Smith[1]
[1]Pacific Northwest National Laboratory Richland, WA 99352
[2]Jet Propulsion Laboratory, Pasadena, CA
[3]University of Southern California, Los Angeles, CA
[4]Argonne National Laboratory, Argonne, IL
margie.romine@pnl.gov

We report on the use of coupled high resolution separation and high mass accuracy and sensitivity Fourier transform ion cyclotron resonance (FTICR) mass spectrometry to characterize the protein content of outer membrane vesicles from the dissimilatory metal-reducing bacterium *Shewanella oneidensis* MR-1. Outer membrane vesicles (MVs) are unique to Gram-negative bacteria, are initiated by the formation of "blebs" in the outer membrane, and are released from the cell surface during growth, trapping some of the underlying periplasmic contents in the process. Membrane vesicles provide an excellent means to identity proteins that are localized to the outer portions of the MR-1 cell envelope without disturbing cellular integrity or the need to further fractionate cells. Mass spectrometric analyses of vesicles isolated from MR-1 cells grown on LB supplemented with fumarate and lactate revealed the presence of 18 outer membrane and 12 periplasmic proteins. Proteins that were identified include electron transport pathway components (OmcA, OmcB, MtrB, CymA, fumarate reductase, and formate dehydrogenase alpha and Fe-S subunits), five putative porins, three proteases, proteins involved in protein maturation (PpiD and DsbA), and two transport proteins (long-chain fatty acids and tungstate). In addition, these samples contained FlaA flagellin proteins and the MshA pilin protein, head and tail proteins from prophage LambdaSo and MuSo2 which, along with several other putative inner membrane and cytoplasmic proteins, probably co-purified with vesicles. The presence of phage coat proteins in these samples suggests that a fraction of cells within MR-1 cultures are undergoing lysis during culture and may explain why proteins predicted to be associated with the inner membrane or cytoplasm were also detected in MV preparations. A comparison of the mass spectrometric results to heme stained gels suggests that at least two additional *c*-type cytochromes are present in the sample, one most likely being MtrA and the second a small *c*-type cytochrome under 14 Kd. We believe that these proteins were not detected by mass spectrometry because of the added mass of heme and are focusing efforts on methods to detect heme modified peptdies in MR-1 *c*-type cytochromes. The presence of electron transport proteins shown *in vitro* to be capable of reducing Fe(III) is consistent with related findings in *S. putrefaciens* CN32, a close relative of MR-1, where vesicles have been shown to mediate Fe(III), U(VI) and Tc(VII) reduction. Future directions will focus on defining the proteome of MR-1 MVs from cells cultured with different electron acceptors.

This work was supported by the Office of Biological and Environmental Research of the U.S. Department of Energy. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute through Contract No. DE-AC06-76RLO

130. *Shewanella Federation*: Data Analysis and Integration

Eugene Kolker
The Institute for Systems Biology, Seattle, WA
ekolker@systemsbiology.org

One of the major objectives of the DOE *Microbial Cell Project* [1] and more generally of the *Genomes to Life Initiative* [2] is "to develop the computational capabilities to integrate and understand genomic, proteomic, metabolic, regulatory, and physiology data and begin to model complex biological systems". This necessitates combining the creativity of interdisciplinary teams incorporating complementary perspectives and principles from diverse fields and across wide strata of academic backgrounds. Work proposed by the *Shewanella Federation* [3] will integrate whole genome experimental approaches, including gene expression arrays and global protein expression studies, with

comprehensive data analysis and modeling as well as biochemical, physiological, and genetic experiments. We will overview data analysis and integration issues associated with the *Shewanella Federation.* We will specifically focus on our recent advances in gene and protein expression studies of model microorganisms and how those advances will be integrated and applied to our study of *Shewanella oneidensis*.

References:

1. Drell, D., *The DOE Microbial Cell Project: A 180° Paradigm Shift for Biology.* OMICS: A Journal of Integrative Biology, 2001, 6(1), in press.
2. http://doegenomestolife.org/
3. http://shewanella.org/

This presentation describes a joint work with S. Stolyar, A. Keller, A. Nesvizhskii, D. Goodlett, E. Yi, S. Purvine (ISB), B. Tjaden, D. Haynor, A. Siegel (UW), A. Smith (UM), C. Rosenow (Affymetrix), E. Koonin (NCBI) as well as other members of the *Shewanella Federation.*
Correspondence to: ekolker@systemsbiology.org, 206.732.1278, fax 206.732.1260.

131. Integrated Analysis of Protein Complexes and Regulatory Networks Involved in Anaerobic Energy Metabolism of *Shewanella oneidensis* MR-1

Jizhong Zhou[1], Frank Larimer[1], James M. Tiedje[2], Kenneth H. Nealson[3], Richard Smith[4], Timothy Palzkill[5], Bernhardt O. Palsson[6], Carol Giometti[7], Dong Xu[1], Mary Lipton[4], Alex S. Beliaev[1], Dorothea K. Thompson[1], Matthew W. Fields[1], James R. Cole[2], and Joel Klappenbach[3]
[1]Oak Ridge National Laboratory
[2]Michigan State University
[3]University of Southern California
[4]Pacific Northwest National Laboratory
[5]Baylor College of Medicine
[6]University of California at San Diego
[7]Argonne National Laboratory
zhouj@ornl.gov

Large-scale sequencing of entire genomes represents a new age in biology, but the greatest challenge is to define gene functions and their regulatory networks at the whole-genome/proteome level. The key goal of this project is to explore whole-genome sequence information for understanding the genetic structure, function, regulatory networks and mechanisms of anaerobic energy metabolism in the metal-reducing bacterium, *Shewanella oneidensis* MR-1. Towards this goal, the following objectives will be achieved: (**1**) We will perform genome-wide mutagenesis using high-throughput random and targeted approaches to understand the functions of genomic sequences with emphasis on anaerobic energy metabolism and environmental responses. (**2**) We will dissect the regulatory networks and mechanisms of the proteins involved in anaerobic energy metabolism using integrated high-throughput genomic, proteomic and bioinformatic approaches by comparing gene expression patterns for both mutant and wild-type cells under different growth conditions. (**3**) We will define the functions of hypothetical proteins involved in anaerobic energy metabolism using integrated bioinformatic, genomic and proteomic approaches together with conventional biochemical methods. (**4**) Finally, we will simulate and predict the metabolic capabilities and cellular dynamics of *S. oneidensis* MR-1 in *silico* using constraints-based modeling approaches. Also, we will construct a central database to efficiently exchange and manage the massive data generated from this project. The proposed project will generate important information about the molecular mechanisms and regulatory networks of anaerobic energy metabolism and environmental responses. The genes identified in this study can be utilized as alternative molecular markers to measure activity and effectiveness of *in situ* bioremediation and will be valuable for the genetic engineering of bacteria for bioremediation purposes.

This research will be conducted as a collaborative project by scientists at Oak Ridge National Laboratory (ORNL), Michigan State University (MSU), University of Southern California (USC), Baylor College of Medicine (BCM), Pacific Northwest National Laboratory (PNNL), Argonne National Laboratory (ANL), and the University of California at San Diego (UCSD).

132. Optical Map Based Sequence Validation of Microbes

Marco Antoniotti[1], Thomas Anantharaman[2], Violet Chang[1], David Schwartz[3], and Bud Mishra[1]
[1]NYU Courant Bioinformatics Laboratory
[2]Biostatistics and Medical Informatics Department, University of Wisconsin
[3]Laboratory for Molecular and Computational Genomic, Departments of Genetics and Chemistry, University of Wisconsin
marcoxa@cs.nyu.edu

The research activity of NYU bioinformatics groups is centered on the algorithms for mapping and sequencing projects and has been focused on the specification languages, environments and systems for bioinformatics. Over the last eight months, we have used the bioinformatics system to develop a suite of mathematical models and associated algorithms for the Validation, Alignment, and Restriction Fragments Translocation Detection and Correction using publicly available optical mapping data and related sequence data. These problems are ideal for testing out the suitability of our software system as it addresses new challenges posed by the availability of vast amount of biological data, especially in the form of DNA sequences. In particular, the Ordered Restriction Mapping bioinformatics manipulation tools, we have developed, serve the dual purpose of improving the DNA sequencing efforts and providing new analysis capabilities that can be derived from the maps themselves directly.

The Validation algorithm we have developed for Ordered Restriction Maps serves to establish the quality of an assembled DNA sequence by comparing it with to an Ordered Restriction Map (e.g. a map obtained via the Optical Mapping Process). The core procedure of the Validation algorithm takes a DNA sequence (retrievable from a variety of sources) and an Ordered Restriction Map (called the "consensus" map). From the DNA sequence, an "in silico" ordered restriction map is obtained (called the "sequence" map). The sequence map is aligned against the original map using a sophisticated dynamic programming formulation. The procedure constructs several alignments of the sequence map against the original map. Each such alignment is ranked according to the value of the computation of a Maximum Likelihood Estimate of a statistical model that takes into account several error sources of the underlying biochemical process. For the Optical Mapping process the error sources considered are

- Sizing errors,
- Missing cuts (cuts appearing in the sequence map and not in the consensus map),
- False cuts (i.e. false positives, cuts appearing in the consensus maps and not in the sequence map).

Orientation is also taken into account, since the consensus map can be in either 3' or 5' sense.

The (Multiple) Alignment algorithm takes as input an Ordered Restriction (Consensus) Map and a set of (small) sequence contigs. Each sequence contig is run through the Validation subsystem and its possible alignments are computed and set aside for future post-processing. Subsequently the Alignment algorithm constructs a putative anchoring of every contig on the consensus map, by selecting one alignment per contig. The selection procedure represents a trade-off among the following criteria

1. Given two contigs alignments, they must not overlap, since if such an overlap were relevant then the sequence assembly algorithm would have already used it.
2. The number of contigs included in the result must be as large as possible,
3. The overall score (obtained from the Validation procedure) must be minimized.

Objective 2 and 3 are contradictory; hence we developed a Lagrangian-like approximation scheme that weighs one or the other criteria under user control. The first criteria may be relaxed in order to look for overlaps that were possibly missed by the contig sequencing and assembly procedure. As our preliminary analysis led us to believe that the problem of construction is likely to be computationally infeasible, we developed two procedures that approximate its construction. The first is a simple Greedy approach that has worked

well in our experiments, and the second is an iterative 1D Dynamic Programming procedure that minimizes a weighted cost function subject the constraints 1 to 3 above.

The Restriction Fragment Translocation (RFT) Detection and Correction algorithm reuses the basic Validation algorithm by considering a consensus map, a sequence map and all the sub-sequences of the sequence map. The validation algorithm is run on the N^2 sub-maps of the sequence map, in order to determine whether any of the sub-maps can be anchored in a different position than the one assigned by the sequence map alignment. The result of the RFT algorithm is an ordered set of rearrangements of the sequence sub-maps.

The three Ordered Restriction Maps algorithms (*Validation, Alignment, and RFT Detection/Correction*) do not work in isolation. While we developed the mathematical and statistical models that constitute the core of the three algorithms, we also developed a software infrastructure integrating the components and based on a DataBase. To achieve the software integration, we developed the specification of file exchange formats and several auxiliary programs used in a variety of ways (e.g. a sequences and maps "simulator" which can be used to generate in silico sequences and maps of various complexity and structure). The three algorithms produce large data sets. In order to navigate the data sets in a more interactive way, we developed two specialized viewers called "CONVex" and "genscape." "genscape" is an evolution of "CONVex." The viewers interact with the underlying infrastructure. The main idea behind the two viewers is to provide a zoomable view of the set of alignment, and to enable the user to inspect the displayed maps (consensus and sequence) at a fine detail level. The viewers are also available as libraries and have been integrated in the VALIS system. All of this software infrastructure has been made available publicly through the Internet and has also been made specifically available to University Wisconsin.

The three algorithms have been tested in a variety of ways. In particular we concentrated on analyzing the *P. falciparum* parasite (the Malaria agent): an organism for which there are both published sequences and published Ordered (Optical) Restriction Maps. We downloaded the known sequences of the *P. falciparum* parasite's 14 Chromosomes from the PlasmoDB online database (www.plasmodb.org). Only Chromosome 2 and 3 have been fully assembled so far. For the remaining 12 we only have sets of contigs, for which no known position along the respective chromosome is published. We ran the Validation algorithm on the *P. falciparum* chromosome 2 and 3 sequence data, against the Ordered (Optical) Restriction Maps. The results show very good agreement between the consensus map and sequence map. Hence we can conclude that both consensus and sequence maps are correct. For the remaining 12 *P. falciparum* chromosomes, we ran the Alignment algorithm and we were able to propose anchoring positions for all the contigs along the respective consensus maps. All these results are viewable at http://bioinformatics.cat.nyu.edu/valis under the "Projects" link.

133. Interaction of Cytochrome c3 with Uranium

Judy D. Wall and Barbara Rapp-Giles
Biochemistry Department, University of Missouri-Columbia
Wallj@missouri.edu

Several years ago, the reduction of soluble U(VI) to U(IV), a much less soluble form, was demonstrated in cell extracts of *Desulfovibrio vulgaris* Hildenborough with hydrogen gas as the reductant. Further experimentation demonstrated that the reduction was dependent on presence of cytochrome c3 in the extract and, of course, hydrogenase. To determine whether cytochrome c3 is the actual reductase of bacteria in this genus, we are preparing purified protein for the application of analytical tools by our Los Alamos collaborators led by Dr. William Woodruff. We have constructed a mutant of *D. desulfuricans* carrying an interrupted cycA gene encoding cytochrome c3. We are now attempting to create a stable mutation by marker exchange mutagenesis. This mutant strain will be necessary to produce mutant forms of the protein for testing the mechanism of the interaction of the protein with uranium, if any.

134. Genome-Wide Functional Analysis of the Metal-Reducing Bacterium *Shewanella oneidensis* MR-1: Progress Summary

Alexander Beliaev[1], **Dorothea K. Thompson**[1,2], Carol S. Giometti[3], Kenneth H. Nealson[4], Alison E. Murray[2], James M. Tiedje[2], and Jizhong Zhou[1,2].
[1]Environmental Sciences Division, Oak Ridge National Laboratory[†], Oak Ridge, TN
[2]Center for Microbial Ecology, Michigan State University, East Lansing, MI
[3]Argonne National Laboratory, Argonne, IL
[4]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA
thompsondk@ornl.gov

Large-scale sequencing of entire microbial genomes has ushered in a new era in biology, but the greatest challenge will be to define gene function and complex regulatory networks at the whole-genome level. In this Microbial Genome Project, we proposed to conduct a microarray-based functional genomic study to elucidate the genes and regulatory mechanisms involved in energy metabolism in *Shewanella oneidensis* MR-1. To study the genes and regulatory schemes underlying anaerobic respiration, wild-type and mutant strains of *S. oneidensis* were examined using DNA microarrays containing 691 open reading frames (ORFs) and 2-D polyacrylamide gel electrophoresis (2-D PAGE). Insertional mutants defective in the Fnr-like *etrA* (electron transport regulator A) and *fur* (ferric uptake regulator) genes were generated by suicide plasmid integration and characterized. Disruption of the *etrA* gene resulted in altered mRNA levels for 69 genes with predicted functions in energy metabolism, transcription regulation, substrate transport, and biosynthesis. In this subset, up to a 12-fold decrease in mRNA abundance was displayed by genes involved in anaerobic respiration (*dmsAB*, *hydABC*, *fdhAC*), while aerobic genes encoding cytochrome oxidases, NADH dehydrogenase, and TCA cycle enzymes were induced up to 3-fold as a result of the *etrA* mutation. Notably, disruption of *etrA* affected the transcription of ten regulatory genes, including *fur* and *hutC* (histidine utilization?). Our results suggest that EtrA plays a subtle role in MR-1 anaerobic gene regulation and is not essential for growth and reduction of electron acceptors.

Microarray analysis of a *fur* knockout strain (FUR1) revealed that genes with predicted functions in electron transport, energy metabolism, transcription regulation, and oxidative stress protection were either repressed (*ccoNQ*, *etrA*, cytochrome *b*- and *c* maturation-encoding genes, *qor*, *yiaY*, *sodB*, *rpoH*, *phoB*, *chvI*) or induced (*yggW*, *pdhC*, *prpC*, *aceE*, *fdhD*, *ppc*) in a *fur*⁻ background. As expected, disruption of *fur* also resulted in derepression of genes putatively involved in siderophore biosynthesis and iron uptake. Analysis of a subset of the FUR1 proteome (i.e., primarily soluble cytoplasmic and periplasmic proteins) indicated that 11 major protein species reproducibly showed significant ($P < 0.05$) differences in abundance relative to the wild type. Protein identification using mass spectrometry revealed that the expression of two of these proteins (SodB and AlcC) correlated with the microarray data. Microarray data and sequence analysis suggest that Fur may act with EtrA and possibly other regulatory proteins to coordinate the synthesis of iron-containing enzymes and cytochromes with iron uptake and respiration. While our findings agree with previous descriptions of Fur as a repressor of iron acquisition genes, this study also suggests that MR-1 Fur plays a role in the coordinate regulation of energy metabolism.

In response to changes in redox and growth conditions, 121 genes out of the 691 arrayed ORFs displayed at least a 2-fold difference in transcript abundance in wild-type *S. oneidensis* MR-1. Genes induced during anaerobic respiration included those involved in cofactor biosynthesis and assembly (*moaACE*, *ccmHF*, *cysG*), substrate transport (*cysUP*, *cysTWA*, *dcuB*), and anaerobic energy metabolism (*dmsAB*, *psrC*, *pshA*, *hyaABC*, *hydABC*). Transcription of genes encoding a periplasmic nitrate reductase (*napDAGHB*), cytochrome c_{552}, and prismane was elevated 8- to 56-fold in response to the presence of nitrate, while *cymA*, *ifcA*, and *frdA* were specifically induced 3- to 8-fold under fumarate-reducing conditions. In addition, we have conducted experiments with *S. oneidensis* MR-1 partial microarrays to determine differential gene expression under iron- and manganese-reducing conditions. Complete linkage hierarchical cluster analysis identified clusters with constitutively expressed genes, those that were anaerobically or aerobically induced with both

Mn(IV) and Fe(III) serving as terminal electron acceptors, and those that were either induced with iron and not manganese or vice versa. Several electron transport carriers including NADH dehydrogenase, ubiquinone (*ubiH*), cytochromes *b* and c_1, and a membrane-bound *c*-type oxidase were induced under all Mn(IV) and Fe(III) experiments. Genes encoding a number of electron transport carriers (dehydrogenases and cytochromes) as well as stress response proteins were induced only with iron as the electron acceptor. Perhaps the most interesting gene, and one with the highest induction under iron reduction, was one encoding N-acylhomoserine lactone synthase, a key regulator of quorum sensing in Gram-negative bacteria. These experiments demonstrate that genes unique to different electron acceptors can be revealed by microarray hybridization. Additional experiments using mutagenesis and whole-genome microarrays (expected to be completed by the end of 2001) will be conducted in order to define the components and mechanisms of metal-reducing pathways in MR-1.

Finally, partial microarrays have been used to define genome relationships in the *Shewanella* genus. DNA:DNA hybridization experiments allowed us to visualize the relationships between organisms in the *Shewanella* genus by comparing individual ORF hybridizations for partial genome arrays. Results from those experiments have shown that other *Shewanella* species hybridize to the MR-1 array, and that some suites of electron accepting and regulatory genes (e.g., *arcA*) are highly conserved within the halotolerant branch of the *Shewanella* genus. Thus, we believe that gene expression data obtained in the proposed work will be relevant to a broader diversity of organisms that are ubiquitous in the environment.

†Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract number DE-AC05-00OR22725.

135. Microarray Analysis of Sugar Metabolism Gene Networks in *Thermotoga maritima*

Arvin D. Ejaz[1], Amy M. Mikula[1], Tu Nguyen[2], Ken Noll[2], Karen E. Nelson[1], and **Steven R. Gill**[1]
[1]The Institute for Genomic Research, Rockville, MD 20850
[2]Department Of Molecular Biology, University of Connecticut, Storrs, CT 06269
srgill@tigr.org

The thermophilic bacterium, *Thermotoga maritima*, is a heterotrophic organism capable of metabolizing complex carbohydrates such as cellulose and xylan. Since cellulose and xylan are major components of plant biomass, their conversion into fuels and chemicals has a significant economic potential. Once the regulation and dynamics of *T. maritima*'s genes are understood, it may become an important organism in high temperature industrial processes that convert plant biomass into useful energy. Previous sequencing of the *T. maritima* MSB8 genome identified genes involved in cellulolytic and xylanolytic pathways. In an effort to elucidate *T. maritima* genes and regulatory networks involved in metabolism of simple and complex carbohydrates, we have constructed a whole genome *T. maritima* microarray. We are currently using these arrays to investigate gene regulation of *T. maritima* grown in continuous cultures with media containing either glucose, lactose or maltose as carbon sources. We will present data on *T. maritima* microarray construction and the results of these experiments.

136. Gene Expression Profiles in *Nitrosomonas europaea*, An Obligate Chemolithoautotroph

Daniel Arp[1], Martin Klotz[2], and Jizhong Zhou[3]
[1]Oregon State University
[2]University of Louisville
[3]Oak Ridge National Laboratory
arpd@bcc.orst.edu

Ammonia-oxidizing bacteria are participants in both the C and N cycles. These bacteria are obligate autotrophs—they obtain all of their carbon for growth from CO_2, and obligate chemolithotrophs—they derive all their reducing power and energy necessary for biosynthesis from the transformation of NH_3 to NO^{2-}. Ammonia-oxidizing bacteria can have profound effects on the environment. Upon oxidation, ammonia applied to croplands is mobilized and can leach into ground and surface waters. Ammonia-oxidizing bacteria also produce the greenhouse gases NO and N_2O. Given the broad substrate specificity of the oxygenase that initiates the oxidation of ammonia, ammonia-oxidizing

bacteria also have the potential to initiate the degradation of several environmental pollutants (e.g. trichloroethylene). Basic and applied research is essential to understand how ammonia-oxidizing bacteria respond to changes in their environment. *Nitrosomonas europaea* is the best characterized of these bacteria. To date, molecular investigations of *N. europaea* have focused primarily on single genes and enzymes involved in the oxidation of ammonia. These studies have revealed surprisingly strong responses to environmental changes, especially given the obligate dependence on ammonia and CO_2. With the sequencing of the genome of this bacterium through the DOE Microbial Genome Program (http://spider.jgi-psf.org/JGI_microbial/html/nitrosomonas_homepage.html), and the development of methods for genetic manipulation, it is now possible to investigate the expression of the entire complement of *N. europaea* genes by applying microarray-based genomic technology. The results will provide insights into how this bacterium responds to changes, but should also provide insights to how autotrophs and lithotrophs in general are modulating their gene expression in response to nutrient changes and environmental stresses.

The specific research objectives of the research are to:

1. Determine the inventory of genes expressed in growing cells and in resting cells of *N. europaea* using high-density whole genome microarrays.
2. Investigate the differences in gene expression for *N. europaea* cells subjected to nutrient shifts, changes in environmental conditions, or environmental pollutants.
3. Identify regulatory genes involved in metabolism and environmental responses through mutagenesis and expression studies.

137. Improving Functional Analysis of Genes Relevant to Environmental Restoration via an Analysis of the Genome of *Geobacter sulfurreducens*

Derek R. Lovley, Madellina Coppi, Stacy Cuifo, Susan Childers, Ching Lean Franz Kaufmann, Daneil Bond, Teena Mehta, and Mary Rothermich
Department of Microbiology, University of Massachusetts, Amherst, MA 01003
dlovley@microbio.umass.edu

Better information is required in order to predict the function of genes, in pure cultures of microorganisms or microbial communities, that are involved in important environmental processes such as the remediation of toxic wastes. *Geobacter* species have novel physiological characteristics that make them ideally suited for the bioremediation of radioactive metals and organic contaminants in subsurface environments. Furthermore, molecular studies have demonstrated that *Geobacter* species are dominant members of microbial communities in geographically and geochemically diverse subsurface environments in which microorganisms are actively bioremediating metal or organic contamination. This represents a rare instance in which an organism that is known to be numerically significant and active in an environmental process of interest is also available in pure culture.

Analysis of the complete genome sequence of *Geobacter sulfurreducens* has revealed that this organism has a high percentage of genes for putative electron-transport proteins and that the expression of these genes is likely to be highly regulated. In order to begin elucidating the function of genes involved in electron transport to metal electron acceptors, *G. sulfurreducens* was grown under steady-state conditions in chemostats with different electron acceptors. Molecular and biochemical analyses demonstrated that genes for a number of c-type cytochromes were specifically expressed only when *G. sulfurreducens* was grown with Fe(III) as the electron acceptor. They were not expressed when fumarate served as the electron acceptor. Physiological studies of knock-out mutants that no longer expressed certain c-type cytochromes suggested that some of the c-type cytochromes are

intermediary electron transport proteins, but that at least one of the c-type cytochromes might function as a terminal metal reductase. Comparison of gene expression between cells grown on soluble Fe(III)-citrate and insoluble Fe(III) oxide demonstrated that genes for the production of pili are specifically expressed during growth on Fe(III) oxide. Differential production of pili was confirmed with electron microscopy. A knock-out mutation that eliminated expression of pilA, the gene for the structural pilin protein, had no effect on the ability of the cells to grow with Fe(III) citrate, but abolished their ability to grow with Fe(III) oxide as the electron acceptor. Complementation with a functional pilA gene restored the capacity for growth on Fe(III) oxide. This is the first description of a protein specifically required for a dissimilatory metal-reducing microorganism to grow on Fe(III) oxide, the primary electron acceptor for metal-reducing microorganisms in subsurface environments.

These results demonstrate that the strategy of examining differential gene expression followed by a genetic analysis of gene function is rapidly elucidating important aspects of *Geobacter* physiology. This in turn, is providing important insights into the mechanisms by which *Geobacter* functions in the bioremediation of metal and organic contamination in the subsurface.

138. Genome Sequencing of *Gemmata obscuriglobus*

Naomi Ward[1], Margaret K. Butler[2], Rebecca L. Smith[2], and John A. Fuerst[2]
[1]The Institute for Genomic Research, Rockville, MD 20850
[2]Department of Microbiology and Parasitology, University of Queensland, Brisbane, Queensland 4072, Australia
nward@tigr.org

Gemmata obscuriglobus is a member of the planctomycete group of Bacteria. These organisms possess a unique combination of morphological and ultrastructural properties, including budding replication, the presence of crateriform structures of unknown function on the cell surface, a diverse range of extracellular appendages, and lack of the "universal" cell wall polymer peptidoglycan. In recent years, the planctomycetes have been found to be widely distributed and often numerically abundant in both aquatic and terrestrial environments. This group also includes the "missing lithotrophs" performing the "anammox" process - organisms that can break down ammonia in wastewater anaerobically; this ecological niche has been postulated for many years, but the organisms performing this role have only recently been identified as planctomycetes. Other proposed environmental roles include degradation of chitin in marine systems, and the breakdown of toxic algal blooms. Lastly, they are a phylogenetically distinct lineage within the Bacteria, and there are currently no published genome sequences from members of this group. *G. obscuriglobus* was the first planctomycete, and indeed first bacterium, shown to possess a membrane-bounded DNA-containing nuclear region, i.e., a structure analogous to the eukaryotic nucleus. This provides an ultrastructural exception to the prokaryote/eukaryote dichotomy, and has interesting implications for transport within the cell, and the linking of transcription and translation processes. These unique features among the Bacteria may have wide implications for discovery of new mechanisms in molecular cell biology correlated with cell compartmentalization. Other planctomycetes were subsequently shown to exhibit various types of cellular compartmentalization, suggesting that this may be a widespread property of members of this group. The genome size of *G. obscuriglobus* is at the upper limits of known bacterial genomes; PFGE-based analyses suggest a genome size of approximately 9Mb. Comparable genome sizes are seen in developmentally complex bacteria such as *Myxococcus xanthus* and *Streptomyces spp.* Availability of genome sequence data from *G. obscuriglobus* will allow comparative analysis of another large genome, and insight into the evolutionary mechanisms which have led to these genomic expansions. At the time of writing, the genome sequencing project was in the library construction phase. A summary of the current status of the project will be presented.

139. Genome Sequence of *Methanococcus maripaludis*, a Genetically Tractable Methanogen

Erik L. Hendrickson[1], Maynard Olson[2], Gary Olsen[3], and **John A. Leigh**[1]
[1]Department of Microbiology, University of Washington
[2]Genome Center, University of Washington
[3]Department of Microbiology, University of Illinois
leighj@u.washington.edu

We have sequenced the genome of *Methanococcus maripaludis* strain LL, a mesophilic methanogenic archaeon, to six-fold coverage. Assembly of the partial sequence has yielded 163 contigs, ranging in size from 0.19 to 106 kb in length covering a total of 1.71 Mb. The total is comparable to that of the most closely related organism with a complete genome sequence, *Methanococcus jannaschii*, with a total genome length of 1.66 Mb. GC content is 33%, again comparable to *M jannaschii* (31%). Potential open reading frames have been identified by the CRITICA program, which predicts 1742 orfs, similar to *M. jannaschii*, with 1738. 1522 of the predicted orfs yield BLAST homologies, the majority of which show their highest homologies to *M. jannaschii* (67%). The rest have their highest homologies to other methanogens (13%), other Archaea (8%), and Bacteria (10%), with only a few having their best match in Eukarya (1%) and four matching viruses. As *M. maripaludis* derives its energy and carbon from formate or H_2 and CO_2, we examined the sequence for the corresponding metabolic genes. The sequence contains genes for a complete methanogenesis pathway as well as formate dehydrogenase, with similar organization to that occurring in *M. jannaschii*. Preparations are under way to finish the genome, and for post-genomic studies.

Collaborators include M. Hackett, R. Bumgarner, and R. Samudrala (University of Washington), W. Whitman and J. Amster (University of Georgia), and D. Söll (Yale University).

140. The Genome of *Ferroplasma acidarmanus*: Clues to Life in Acid

Larry Croft[1], Amanda Barry[2], Paul Predki[3], Stephanie Stilwagen[3], Genevieve Johnson[3], Thomas M. Gihring[1], Brett J. Baker[1], Jennifer Macalady[1], George F. Mayhew[4], Valerie Burland[4], Teresa Janecki[3], Charles W. Kaspar[5], Brian Fox[2], and **Jillian F. Banfield**[1]
[1]Department of Geology and Geophysics, University of Wisconsin-Madison
[2]Biochemistry Department, University of Wisconsin-Madison
[3]DOE Joint Genome Institute
[4]Genome Center, University of Wisconsin-Madison
[5]Department of Food Microbiology and Toxicology, University of Wisconsin-Madison
jill@seismo.berkeley.edu

The archaeon *Ferroplasma acidarmanus* populates hot, acidic (pH 0-3), metal-rich solutions associated with acid mine drainage sites. The multiple challenges posed by its environment make it an excellent model organism for study of genes involved in metal and acid tolerance. Because it is an obligate acidophile, it is also ideal for investigation of lateral gene transfer as it is effectively separated from most of the biosphere by its environment. The 2Mb genome was sequenced and is 15% homologous (blast expectation < 1e-45) by peptide similarity to *T. acidophilum*, an acidophilic scavenger, the closest sequenced relative to *F. acidarmanus*. Incomplete amino acid biosynthetic pathways and an array of intra and extracellular proteases and amino acid pumps support heterotrophic growth and indicate a complex dependence on other community members for essential organic compounds.

The *F. acidarmanus* proteome is also 4% homologous to *Sulfolobus solfataricus*, another archaeal acidophile. This is much more than would be expected by evolutionary relatedness alone. It is highly likely that lateral gene transfer has occurred between *Sulfolobus* spp. and *F. acidarmanus*. Bacteriophage sequences and transposons in the genome suggest vehicles for lateral gene transfer.

Using TMHMM, 23% of proteins in the *F. acidarmanus* proteome were identified as membrane bound and 12% of proteins were permeases or created permease-like structures. Over 25% of all proteins have not been previously identified (singletons) and a large proportion (39%) of these are membrane bound, suggesting much unknown membrane associated extracellular activity. 23% of proteins have similarity to proteins from other organisms but have unknown function, most of these are also membrane bound (18% of proteome). There appears to be no discernible amino acid bias between cytoplasmic peptides and regions of peptides exposed to the extracellular environment. This suggests that protein secondary or tertiary structure or modification plays a significant role in acid stability of extracellular proteins.

Acid mine drainage contains high concentrations of toxic metal species such as arsenic, which are removed from the cytoplasm by a set of metal efflux pumps, and detoxified by metal reducing enzymes (such as mercury reductases). Further protection is afforded by a predominantly tetraether-linked lipid membrane monolayer that makes feasible the large proton gradient between the pH ~ 5.2 cytoplasm and the pH < 1.0 environment.

Several *F. acidarmanus* genes have been expressed in *E. coli*, including two Rieske-type iron-sulfur proteins, a blue copper protein, two cytochrome p450-like proteins, a cobalamin biosynthesis protein, and an iron superoxide dismutase. Proteomic studies of gene expression when *F. acidarmanus* is grown heterotrophically and on ferrous iron are currently underway. Protein expression studies of a candidate tetraether lipid synthesis pathway are in progress.

141. Genome Sequence of the Metal-Reducing Bacterium, *Shewanella oneidensis*

John F. Heidelberg[1], Ian T. Paulsen[1], Karen E. Nelson[1], William C. Nelson[1], Jonathan A. Eisen[1], Barbara Methe[1], Eric J. Gaidos[3], Owen White[1], Kenneth H. Nealson[2], and Claire M. Fraser[1]
[1]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850 USA
[2]University of Southern California
[3]University of Hawaii
jheidel@tigr.org

The *Shewanella oneidensis* genomic sequence will expedite efforts to use this organism for bioremediation of dissolved toxic metals and organic toxins from water supplies. This bacterium, and other metal-reducing bacteria, uses metals (rather than oxygen) as the terminal electron acceptors for anaerobic respiration. This respiratory capability makes it a valuable tool in the removal of toxic metals such as uranium and chromium. Here we report the compete genome of *S. oneidensis* MR-1. *Shewanella oneidensis* is notable for its inability to grow on a wide variety of carbon sources, rather, the organism seems to be tuned to the use of fermentation end-products, and the use of the pyruvate formate lyase reaction under anaerobic conditions. *Shewanella oneidensis* is unusual for a gamma proteobacterium, containing a very high number of multi-heme cytochrome c genes, and having the diversity of electron transport capacities it possesses. Insertion sequences compose 5.5% of the genome sequence and, likely play a critical role in shaping the genome.

142. The *Colwellia* Strain 34H Genome Sequencing Project

Barbara Methe[1], Matthew Lewis[1], Bruce Weaver[1], Jan Weidman[1], William Nelson[1], Adrienne Huston[2], Jody Deming[2], and Claire Fraser[1]
[1]The Institute for Genomic Research, Rockville, MD 20850
[2]School of Oceanography, University of Washington, Seattle, WA 98195
bmethe@tigr.org

Approximately 7% percent of the Earth's surface is covered by sea ice and by volume about 90% of the world's oceans exist at a temperature of 5°C or less. As a result, large regions of the marine ecosystem are permanently cold and colonized principally by cold-adapted microorganisms. The sequencing of the entire genome of *Colwellia* strain 34H will provide the first complete assembly of a psychrophilic (growth optima < 15°C and maximum growth < 20°C) bacterium. As a member of the gamma sub-class of the proteobacteria, the genus *Colwellia* represents a group of obligate marine bacteria many of which are psychrophiles, that play important roles in carbon and nutrient cycling in polar marine environments. Of particular interest is the capability

of *Colwellia* to produce cold-adapted enzymes that have potentially important applications for use in the fields of biotechnology and bioremediation. Recent biochemical investigations of strain 34H have demonstrated its ability to release cold-adapted extracellular proteases with the lowest activity optima yet reported for a cell-free extract from a pure culture.

With primary funding from the Department of Energy, the *Colwellia* strain 34H genome project has commenced using a random shotgun approach which has already provided approximately eight-fold coverage of a 5.3 Mb genome. Closure of the remaining physical and sequencing gaps and resolving and ordering of RNA operons is now in progress using a variety of directed sequencing strategies including: sequencing from primers designed to point into gaps, multiplex PCR, micro-library construction and transposon mutagenesis of appropriate library clones. A suite of software programs developed by The Institute for Genomic Research is also being employed to assemble the genome, aid in gap closing, assembly verification and annotation. Examination of this genome will improve our understanding of the adaptations of this organism to cold marine environments, which in turn has important implications in areas as diverse as microbial ecology, evolution, biotechnology and bioremediation.

143. Complete Genome Sequence of *Acidithiobacillus ferrooxidans* Strain ATCC23270

Herve Tettelin[1], Keita Geer[1], Jessica Vamathevan[1], Florenta Riggs[1], Joel Malek[1], Maureen Levins[1], Mobolanle Ayodeji[1], Sofiya Shatsman[1], Getahun Tsegaye[1], Stephanie McGann[1], Robert J. Dodson[1], Robert Blake[2], and Claire Fraser[1]
[1]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850
[2]Xavier University, College of Pharmacy, 7325 Palmetto Street, New Orleans, LA 70125
tettelin@tigr.org

Acidithiobacillus ferrooxidans is a Gram-negative bacterium of industrial and environmental relevance. It is a major component in the consortia of microorganisms used in biomining and a contributor to acid mine runoff, which results in pollution near metal and coal mines and other related environments. *A. ferrooxidans* is acidophile: optimal pH between 1.5 and 2.0, mesophile: under 45°C, and chemolithoautotroph. It gains energy from oxidative phosphorylation, obtains nitrogen from N_2 in the air and carbon exclusively from CO_2 fixation. It derives energy from oxidation of reduced inorganic sulfur to H_2SO_4 and oxidation of $Fe2^+$ to $Fe3^+$ which precipitates as insoluble $Fe(OH)_3$. The 2.9 Mb chromosome of *A. ferrooxidans* ATCC23270 was sequenced to 8x coverage with 50,136 shotgun sequences derived from small (ca. 2 kb, 5x) and large (ca. 10 kb, 3x) insert shotgun libraries, and currently undergoes the final stages of gap closure. The genomic sequence displays a G+C content of 58.4% and contains 61 repeats larger than 500 bp. It was annotated in a fully automated fashion, which will be followed by manual curation (expected to occur after this workshop) of each open reading frame. However, the available annotation will be sufficient to derive meaningful information about the metabolic pathways that are critical to the life of this organism in its inorganic environment. In addition, we will focus on the set of surface (including transporters) and secreted proteins that allow *A. ferrooxidans* to feed on ores. Results from analyses on the structure of the chromosome, including regions of atypical nucleotide composition, putative islands of horizontal transfer, recent gene duplications, etc. will also be discussed.

144. The Complete Genome Sequence of the Green Sulfur Bacterium *Chlorobium tepidum*

Jonathan A. Eisen[1], Karen E. Nelson[1], Ian T. Paulsen[1], John F. Heidelberg[1], Martin Wu[1], Robert J. Dodson[1], Robert Deboy[1], Michelle L. Gwinn[1], William C. Nelson[1], Daniel H. Haft[1], Erin K. Hickey[1], Jeremy D. Peterson[1], A. Scott Durkin[1], James L. Kolonay[1], Fan Yang[1], Ingeborg Holt[1], Lowell A. Umayam[1], Tanya Mason[1], Michael Brenner[1], Terrance P. Shea[1], Debbie Parksey[1], Tamara V. Feldblyum[1], Cheryl L. Hansen[1], M. Brook Craven[1], Diana Radune[1], Jessica Vamathevan[1], Hoda Khouri[1], Owen White[1], J. Craig Venter[5], Tanja M. Gruber[4], Karen A. Ketchum[5],

Hervé Tettelin[1], Donald A. Bryant[3], and Claire M. Fraser[1,2]
[1]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850 USA
[2]George Washington University Medical Center, 2300 Eye Street NW, Washington, DC 20037 USA
[3]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA, USA
[4]Department of Stomatology, Microbiology and Immunology, University of California at San Francisco, San Francisco, CA, USA
[5]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850 USA
jeisen@tigr.org

The complete genome of the green-sulfur eubacterium *Chlorobium tepidum* TLS was determined to be a single circular chromosome of 2,154,946 base pairs. This represents the first genome sequence from the phylum *Chlorobia*, whose members perform anoxygenic photosynthesis by the reductive TCA cycle. Genome comparisons have identified genes in *C. tepidum* that are highly conserved among photosynthetic species. Many of these have no assigned function and may play novel roles in photosynthesis or photobiology. Numerous duplications of genes involved in biosynthetic pathways for photosynthesis and the metabolism of sulfur and nitrogen were identified. Thirty-eight percent of predicted proteins with likely roles in central intermediary metabolism are most similar to proteins from Archaeal species. Evidence suggests many of these were acquired by lateral gene transfer.

145. The Genome Sequences of *Bacillus anthracis* Strain Ames

T. D. Read, E. Holtzapple, and S. Peterson
The Institute for Genomic Research
tread@tigr.org

Whole-genome sequencing of a *Bacillus anthracis* Ames isolate (pXO1- pXO2-) is nearing completion. The initial phase of the project, random sequencing of small- and large- insert libraries has been completed and efforts are being directed currently to closing gaps between assemblies. Many portions of the *B. anthracis* sequence appear to have similar gene content and organization to the archetypal non-pathogenic *B. subtilis* and to the recently sequenced *B. halodurans* genome. At least 60% of *B. anthracis* ORFs have homologues to known *B. subtilis* genes. These include many spore-coat and spore-germination determinants believed to play in important role in virulence. There are many genes without homologues in *B. subtilis* that could be important in anthrax infection, including several hemolysins and phospholipase genes. Also notable was the presence in the genome of numerous copies of a conserved 16 bp palidrome known to be a target of the *B. thuriengiensis* positive regulator of extracellular virulence determinants, PlcR. However, the *B. anthracis* plcR gene contains a potential loss-of-function deletion. The pXO plasmids that contain the key virulence genes encoding toxin and capsule have recently been sequenced. Although the plasmids appear to have undergone frequent rearrangements, there are few apparent instances of gene transfer between plasmid and chromosome, suggesting possible recent arrival of the episome into *B. anthracis*. We have also been recently funded to sequence the Ames strain isolated from the Florida bioterror attack and will present strain comparisons with the 'laboratory' Ames strain.

146. The Complete Genome Sequence of *Pseudomonas putida* KT 2440

Karen E. Nelson[1], Burkhard Tuemmler[2], and Claire M. Fraser[1]
[1]The Institute for Genomic Research, Rockville, MD 20850
[2]Medizinische Hochschule Hannover (MHH), Hannover, Germany
kenelson@tigr.org

Pseudomonas putida is a commonly found soil bacterium that is also a biocontrol agent for plant pathogens, and has a broad capacity for bioremediation and biotransformation. In an attempt to characterize the species, the type strain KT2440 was sequenced by the random shotgun procedure. The 6.18 Mb genome is composed of 5427 open reading frames, 11% of which are unique to the bacterium. The genome sequence reveals numerous transport and metabolic systems that relate to the organisms versatility. As expected, there is a high level of gene conservation with the pathogenic species *Pseudomonas aeruginosa*, a major cause of opportunistic human infections including cystic fibrosis. The genomic differences that contribute to

variations in abilities of the two species have been highlighted by whole genome comparisons, and will be presented.

147. Genome Sequence of *Methylococcus capsulatus*

Naomi Ward[1], Jonathan Eisen[1], Claire Fraser[1], George Dimitrov[1], Scott Durkin[1], Lingxia Jiang[1], Hoda Khouri[1], Katherine Lee[1], David Scanlan[1], Nils Kåre Birkeland[2], Live Bruseth[2], Ingvar Eidhammer[2], Svenn H. Grindhaug[2], Ingeborg Holt[2], Harald B. Jensen[2], Inge Jonasen[2], Øivind Larsen[2], and Johan Lillehaug[2]
[1]The Institute for Genomic Research, Rockville, MD
[2]The University of Bergen, Norway
nward@tigr.org

Methylococcus capsulatus (Bath) is a Gram-negative aerobic bacterium (family *Methylococcaceae*, gammaproteobacteria) capable of using methane as a sole carbon and energy source. Methane is oxidized via methanol to formaldehyde, which is either assimilated into cellular biomass or dissimilated to carbon dioxide. Methanotrophs such as *M. capsulatus* are responsible for the oxidation of methane produced through methanogenesis, and are therefore of environmental importance in reducing the amount of greenhouse gases formed in the Earth's atmosphere. *M. capsulatus* (Bath) also has considerable potential for large-scale commercial production of microbial proteins by fermentation, due to its ability to grow to high cell density with only natural gas as a carbon source. The 3.3 Mbp *M. capsulatus* (Bath) genome was sequenced by the random shotgun sequencing strategy, in a collaboration between TIGR and The University of Bergen. At the time of writing, the genome is in gap closure and consists of a single group of contigs assembled from 41,368 individual sequences. A summary of the current status and preliminary annotation will be presented.

148. Comparative Genomic Sequence Analysis of Three Strains of the Plant Pathogen, *Xylella fastidiosa*

S. Stilwagen[1], P. F. Predki[1], A. Bhattacharyya[3], H. Feil[4], W. S. Feil[4], F. Larimer[2], K. Frankel[1], S. Lucas[1], D. Rokhsar[1], E. Branscomb[1], and T. Hawkins[1]
[1]U.S. DOE Joint Genome Institute, Walnut Creek, CA 94598
[2]Oak Ridge National Laboratory, Oak Ridge, TN
[3]Integrated Genomics, Inc, Chicago, IL
[4]Department of Environment, Science, Policy, and Management, University of California, Berkeley, CA
stilwagen1@llnl.gov

The Joint Genome Institute (JGI) has shotgun sequenced the genomes of two strains of the fastidious, xylem-limited bacteria, *Xylella fastidiosa*, to high draft (eightfold coverage). This gram negative bacterium causes a range of economically important diseases which include Pierce's disease (PD) in grapevines and citrus variegated chlorosis (CVC) in citrus plants. The diseases caused by this plant pathogen are responsible for major economic and crop losses globally. We present here the comparative analysis of the ordered and oriented genome sequences of the strains *X.fastidiosa pv. almond* and *X.fastidiosa pv. oleander* versus the finished genome of *Xylella fastidiosa pv. citrus*. Our analyses will illustrate not only the utility of high draft genome sequences but will also identify the signature features of the *Xylella* genomes and reveal the high, yet broad conservation of the gene repertoire across the three strains. We will further present our findings regarding putative candidate genes which have resulted from horizontal gene transfer.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, the Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098, and the Los Alamos National Laboratory under contract No. W-7405-ENG-36.

149. Finishing/Investigating the Genomes of *Prochlorococcus*, *Synechococcus*, and *Nitrosomonas*: An Overview

P. Chain[1], W. Regala[1], L. Vergez[1], S. Stilwagen[2], F. Larimer[3], D. Arp[4], N. Hommes[4], A. Hooper[5], S. Chisholm[6], G. Rocap[7], B. Brahamsha[8], B. Palenik[8], and J. Lamerdin[1]
[1]Lawrence Livermore National Laboratory, Livermore, CA
[2]Joint Genome Institute, Walnut Creek, CA
[3]Oak Ridge National Laboratory, Oak Ridge, TN
[4]Botany and Plant Pathology Department, Oregon State University, Corvallis, OR
[5]Department of Biochemistry, University of Minnesota, St. Paul, MN
[6]Departments of Civil and Environmental Engineering and Biology, Massachusetts Institute of Technology, Cambridge, MA
[7]School of Oceanography, University of Washington, Seattle, WA
[8]Scripts Institution of Oceanography, University of California San Diego, San Diego, CA
chain2@llnl.gov

The output of sequence data from sequencing centers, such as the DOE's Joint Genome Institute, has been rising at an exponential rate for the past decade or two. The increase in sequencing efficiency over the past few years has resulted in a bottleneck shift, from the accumulation of raw data to the finishing, annotation and analysis of genomes. The first two publications describing complete microbial genomes were reported in 1995. Only seven years later, there are approximately 60 complete, annotated microbial genomes available, along with published draft analyses of several multi-cellular eukaryotes. However, an even greater number of projects are either currently underway or are awaiting the finishing process, which provides a complete picture of the genome including contextual information, captures all the sequences missed in the draft phase, and adds a level of confidence to the genomic sequence.

In support of the DOE's Carbon Sequestration and Management Program, we undertook the challenging task of finishing the genomes of three autotrophic bacteria which play unique roles in their soil and ocean ecosystems. The genomes of *Prochlorococcus marinus* MIT9313 and *Synechococcus sp.* WH8103, two cyanobacteria, had been drafted to 7-fold coverage by the JGI, while *Nitrosomonas europaea* sp. Schmidt was at near 14-fold coverage. *Nitrosomonas europaea* is an obligate ammonia-oxidizing beta-proteobacteria that can meet its carbon requirements entirely through the fixation of carbon dioxide, while *Prochlorococcus* and *Synechococcus* are the dominant photosynthetic organisms in the open ocean, contributing to a significant proportion of the earth's biomass. Despite the excess sequence coverage of the *Nitrosomonas*, several genomic structural features made circularization a great deal more difficult than for the two cyanobacterial genomes. With these finished genomes, complete annotation and analysis (including comparative analysis) may help elucidate the pathways relevant to understanding the physiological and genetic controls of photosynthesis, nitrogen fixation and carbon cycling.

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

150. Cloning, Expression, Purification and Initial Characterization of a Three–Heme Cytochrome from *Geobacter sulfurreducens*

Yuri Y. Londer, P. Raj Pokkuluri, William C. Long, and Marianne Schiffer
Biosciences Division, Argonne National Laboratory, Argonne, IL 60439
londer@anl.gov

Multiheme cytochrome c proteins have been shown to exhibit a metal reductase activity, which is of great environmental interest, especially in bioremediation of contaminated sites. *Geobacter sulfurreducens* is one of a family of microorganisms that oxidize organic compounds using Fe(III) or other metals as terminal electron acceptors. We cloned a gene encoding a three-heme 9.6 kDa cytochrome from *G. sulfurreducens* believed to be involved in metal reduction (1) and expressed it in *E. coli* together with cytochrome c maturation gene cluster ccmABCDEFGH on a separate plasmid (2). We designed two different expression systems for the expression and correct post-translational

processing, under control of T7 and lac promoters. We found that N-terminal His-tag is detrimental for proper maturation, where all three hemes are incorporated into the protein. We also established a method for purification of the mature form and species with fewer heme groups. The pure protein has the same molecular weight and displays the same spectra, both in reduced and in oxidized forms, as the protein isolated from *G. sulfurreducens*. Crystals of the recombinant protein were obtained and initial structure determination is under way. This work is a part of an ongoing collaboration with Prof. D. R. Lovley's group at University of Massachusetts.

151. Microbial Metal and Metalloid Metabolism and Beyond

Lynda B. M. Ellis, Larry P. Wackett, Wenjun Kang, Bo Hou, and Tony Dodge
University of Minnesota
lynda@tc.umn.edu

Microbial functional genomics is faced with an ever-growing list of genes that are labeled "unknown" due to lack of knowledge about their function. The majority of microbial genes encode enzymes. Enzymes are the catalysts of metabolism: catabolism, anabolism, stress responses, and many other cell functions. A major problem facing microbial functional genomics is the wide breadth of microbial metabolism, much of which remains undiscovered. The breadth of microbial metabolism has been surveyed by the PIs and represented according to reaction types on the University of Minnesota Biocatalysis/ Biodegradation Database (UM-BBD): http://umbbd.ahc.umn.edu/search/FuncGrps.html

The database depicts metabolism of 50 chemical functional groups, representing most current knowledge. At least twice that number might be metabolized by microbes. Thus, 50% of the unique biochemical reactions catalyzed by microbes could remain undiscovered. Many genes with unknown function, including conserved hypothetical genes, encode functions yet undiscovered. This gap will be partly filled by the current project. The UM-BBD will be greatly expanded as a resource for microbial functional genomics, adding information on biotransformations of metals, metalloids and metal chelators and toxic organics. Two relevant lists are all present UM-BBD pathways: http://umbbd.ahc.umn.edu/servlets/pageservlet?ptype=allpathways and all present UM-BBD metal, metalloid and metal chelator pathways: http://umbbd.ahc.umn.edu/metals.html

This project was initiated with a meeting of its International Advisory Board in late September, 2001. This productive meeting was the start of several important future collaborations on computational and experimental work.

Computational methods will be developed to predict microbial metabolism that is not yet discovered. A concentrated effort to discover new microbial metabolism will be conducted, focused on metabolism of direct interest to DOE: the transformation of metals, metalloids, organometallics and toxic organics; precisely the type of metabolism that has been characterized most poorly to date. These studies will directly impact functional genomic analysis of DOE-relevant genomes.

152. A Potential *Thermobifida fusca* Xyloglucan Degrading Operon

Diana Irwin[1], Mark Cheng[1], Bosong Xiang[2], and **David B. Wilson**[1]
[1]Molecular Biology and Genetics and [2]Chemistry and Chemical Biology, Cornell University
di12@cornell.edu

The annotated genome of *Thermobifida fusca* contains eight potential cellulase genes. Six of these had been previously cloned and sequenced in our laboratory. We subcloned one of the additional genes, contig 40-gene 27, which encoded a glycosyl hydrolase family 74 catalytic domain followed by a family II cellulose binding domain. The expressed and purified protein had low activity on carboxymethyl cellulose and amorphous cellulose, but high activity on xyloglucan. The adjacent upstream gene, contig 40-gene26, encodes a potential alpha-xylosidase gene, suggesting that this region contains a xyloglucan degrading operon. It is

interesting that the gene for Cel9B is close by and it is the only other cellulase that has activity on xyloglucan. Time dependent NMR studies of the products of Cel74A hydrolysis showed that this enzyme uses an inverting mechanism, which would be expected to be used by all family 74 enzymes.

153. Proteome Flux in Photosynthesis and Respiration Mutants of *Synechocystis sp. PCC 6803*

Julian P. Whitelegge[1], Kym F. Faull[1], Robby Roberson[2], and Wim Vermaas[3]
[1]The Pasarow Mass Spectrometry Laboratory, Departments of Psychiatry and Biobehavioral Sciences, Chemistry and Biochemistry and the Neuropsychiatric Institute, UCLA
[2]Department of Plant Biology, Arizona State University
[3]Department of Plant Biology, and Center for the Study of Early Events in Photosynthesis, Arizona State University
jpw@chem.ucla.edu

The availability of complete genome data delivers the potential to identify isolated proteins based upon coincidence of experimental mass data from fragments of polypeptide chain with hypothetical datasets calculated based upon translations of genomic sequences. In order to understand the interaction and control networks of proteins involved in photosynthesis and respiration in the cyanobacterium *Synechocystis sp. PCC 6803*, we are measuring changes in protein expression in populations of cells placed under specific experimental treatments. Early experiments are focusing upon mutants where either Photosystems 1 or 2 are completely knocked out. Two different strategies are being employed in order to fully characterize changes in the proteome. Firstly, 2D-electrophoresis provides a simple way to visualize many of the more abundant proteins of the cell and fluxes of abundance, as well as post-translational modifications that alter mobility in either isoelectric focusing or SDS-PAGE. Secondly, intact protein mass profiles generated by liquid chromatography – mass spectrometry (LC-MS) are used to define the native covalent state of a gene product and heterogeneity associated with it. Moreover, this latter option provides the ability to monitor subtle covalent modifications that are undetectable in 2D-gel systems providing a valuable alternative technology for proteomics. Subfractionation techniques will be applied to monitor less abundant members of the proteome and integrate with parallel studies of ultrastructure and metabolism.

154. Modification of the IrrE Protein Sensitizes *Deinococcus radiodurans* R1 to the Lethal Effects of UV and Ionizing Radiation

Ashlee M. Earl and John R. Battista
Department of Biological Sciences, Louisiana State University and A & M College, Baton Rouge, LA 70803
aearl@lsu.edu

IRS24 is a strain of *Deinococcus radiodurans* carrying mutations in two loci, *uvrA* and *irrE*, rendering it sensitive to the lethal effects of UV and ionizing radiation. These sensitivities can be reversed by introducing the wild type *irrE* allele back into IRS24 via natural transformation. The mutation was localized to a 970bp region containing one putative open reading frame (ORF), DR0167, and 179bp of sequence upstream. Subsequent sequence analysis of the *irrE* allele in IRS24 revealed a transition mutation at codon 111 of DR0167 resulting in an arginine to cysteine amino acid substitution. DR0167 was also inactivated by transposon mutagenesis in the wild type strain, R1. The insertion mutant has a more pronounced sensitivity to both UV and ionizing radiation suggesting that the point mutant has some activity. Blast search analysis of DR0167 reveals only minimal similarity to proteins currently available in the databases. A "weak" helix-turn-helix (HTH) motif was identified within the protein that may indicate a capacity to bind DNA and, perhaps, a potential role for IrrE in gene regulation. In order to test whether the mutation in DR0167 causes a regulatory deficiency we examined the pattern of transcription after applying ionizing radiation, comparing the *irrE* mutant and its parent using DNA microarray technology.

155. The Genome of a White Rot Fungus: How to Eat Dead Wood

Nicholas Putnam[1,2], Jarrod Chapman[1,2], Susan Lucas[1], Luis Larrondo[3], Maarten Gelpke[1,2], Kevin Helfenbein[1], Jeff Boore[1], Randy Berka[4], Doug Hyatt[5], Frank Larimer[5], Dan Cullen[3], Paul Predki[1], Trevor Hawkins[1], and **Dan Rokhsar**[1,2]
[1]U.S DOE Joint Genome Institute, Walnut Creek, CA 94598
[2]University of California, Berkeley CA
[3]Forest Products Laboratory, Madison WI
[4]NovoEnzymes Biotech, Davis CA
[5]Oak Ridge National Laboratory, Oak Ridge TN
dsrokhsar@lbl.gov

White rot fungi produce a suite of unique extracellular oxidative enzymes that degrade lignin, a complex aromatic polymer that is a major component of wood, as well as related compounds found in explosive contaminated materials, pesticides, and toxic wastes. To elucidate the genomic toolkit of these fungi, we have sequenced the thirty million base-pair genome of *Phanerochaete chrysosporium* to high draft using a whole genome shotgun method, making it the first basidiomycete to be sequenced. Assembly of the sequence fragments was carried out using a newly developed algorithm that self-consistently incorporates paired-end information and provides a suite of analysis tools for large scale assemblies. We present an analysis of the *P. chrysosporium* genome, including the major families of secreted enzymes that characterize the white rot fungi, analysis of its mitochondrial genome, and phylogenetic comparisons with more distantly related fungi, animals, and plants.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, the Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098, and the Los Alamos National Laboratory under contract No. W-7405-ENG-36.

156. Metabolic Pathway Elucidation for Microbial Genomes

Imran Shah, Ronald Taylor, and Shilpa Rao
University of Colorado School of Medicine
imran.shah@uchsc.edu

The goal of this work is to develop predictive computational tools for elucidating microbial metabolic pathways. Metabolic inference is becoming increasingly feasible with the availability of large amounts biochemical data. Our system consists of three main modules: (i) a biochemical knowledgebase that integrates data from molecules to pathways and supports deductive inference, (ii) a predictive tool that aids in automated assignment of catalytic functions to putative proteins, and (iii) a pathway synthesis algorithm that generates pathways from catalytic function assignments. We are using this system to analyze the metabolic pathways using whole microbial genomic data.

157. Annotation of *Shewanella oneidensis* MR-1 from a Metabolic and Protein-Family View

Monica Riley and Margrethe H. Serres
Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543
mriley@mbl.edu

Since *Shewanella oneidensis MR-1* (formerly *Shewanella putrefaciens MR-1*) first was isolated in 1988, experimental biochemical studies have been aimed at understanding its interesting energy metabolism and its ability to use metal ions as both electron donors and acceptors. The genome of *Shewanella oneidensis MR-1* has recently been sequenced by TIGR opening the door for full genomic analysis. We are starting the work of determining the complete metabolic and energy transfer capabilities in *Shewanella oneidensis MR-1*. Sequence similar proteins to *E. coli* K-12 and over 40 other genomes are identified using DARWIN analysis. EcoCyc and MetaCyc are initial sources for metabolic pathways. Based on amino acid

alignments of at least 83 amino acids, 57% of the *Shewanella* proteins have sequence similar matches to *E. coli* at a similarity distance of <200 PAM units. Putative functions can be assigned to 51% of the proteins based on matches to *E. coli* alone. Initial analysis of the metabolic pathways shows that *Shewanella* contains sequence similar proteins to a majority of *E. coli* proteins involved in energy metabolism, building block biosynthesis, and intermediate metabolism, but some functions are more closely related to those of other organisms. Details will be presented. Proteins of *Shewanella oneidensis MR-1* are grouped into sequence related groups representing paralogous proteins within the *Shewanella* genome, presumed to have arisen through duplication and divergence either in the *Shewanella* genome or in its ancestors. These groups provide a source for annotation of gene function as well as studying evolution of functions in the organism. Structural predictions for the encoded proteins are in process and will be used in the annotation procedure as well as in protein family studies.

158. Modeling DNA repair in *Deinococcus radiodurans*

Shwetal S. Patel and **Jeremy S. Edwards**
Chemical Engineering, University of Delaware
edwards@che.udel.edu

We are developing novel computational tools to analyze the DNA repair capabilities of *Deinococcus radiodurans* and their relationship to the metabolic capabilities of this organism. Such tools will be extremely useful in providing the necessary insights required to metabolically engineer *D. radiodurans* strains capable of growing under nutrient poor conditions and yet possessing extraordinary DNA repair capabilities. We are moving towards this ultimate goal along two directions. The first involves the construction of a database for the automated construction of metabolic flux balance models. We will then apply flux balance analysis to study the metabolic capabilities of *D. radiodurans* and identify the optimal growth characteristics under different conditions. Additionally, we will study the metabolic pathway structure of *D. radiodurans* to comprehensively examine the metabolic repertoire of *D. radiodurans* and elucidate the regulatory structure of *D. radiodurans*. The second is concerned with the development of dynamic models of the known DNA repair pathways in *D. radiodurans*. The structure of these dynamic models will evolve through critical tests of the hypothesis that the observed DNA repair capabilities of *D. radiodurans* are due solely to known mechanisms. These dynamic models will be used to compute the metabolic flux requirements during DNA repair. This will provide the critical link between the metabolic and DNA repair capabilities in *D. radiodurans*.

In this presentation, we will discuss the current state of our work. In particular, we will discuss a mathematical model to describe the pathway for nucleotide excision repair. Taken together, our analysis will provide valuable information for the metabolic engineering of *D. radiodurans* strains for bioremediation, and our work will significantly contribute to the growing fields of bioinformatics, computational biology, functional genomics and DNA repair.

159. A Novel Combinatorial Biology Method to Functionally Characterize Microbial ORFs

Diane J. Rodi and **Lee Makowski**
Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439
lmakowski@anl.gov

This project applies a novel approach to genome-wide identification of small molecule binding proteins. Preliminary results demonstrated that the similarity between the sequence of a protein and the sequences of affinity-selected, phage-displayed peptides are predictive for protein binding to a small molecule ligand. Affinity-selected peptides provide information analogous to that of a consensus-binding sequence, and can be used to identify ligand-binding sites. Libraries of phage-displayed peptides are being screened for affinity to common metabolites and other small molecules with the goal of applying the affinity-selected sequences to genome-wide identification of proteins that have a high probability of binding to the screened ligands. Our initial experiments have involved affinity selection of ATP-binding peptides. Details of the selection process have been analyzed through the use of 4 different sets of experimental conditions in

order to optimize selection. A comprehensive analysis of the sequences of peptides that contact ATP in ATP-binding proteins whose three-dimensional structures are known has been carried out to provide a basis for analysis of the ATP-selected peptides. Detailed informatic analysis has been used to identify significant and informative differences between the sequences of ATP-binding peptides and the sequences of peptides that contact ATP in ATP-binding proteins. A comprehensive analysis of these sequences is providing insights into the process of molecular recognition and the way ATP interacts with proteins.

160. Annotation of Draft Microbial Genomes

Frank W. Larimer, Loren Hauser, Miriam Land, Doug Hyatt, Manesh Shah, Philip LoCascio, Edward C. Uberbacher, and Inna Vokler
Oak Ridge National Laboratory
fwl@ornl.gov

A draft analysis pipeline has been constructed to provide annotation for the microbial sequencing projects being carried out at the Joint Genome Institute. The pipeline was applied to annotating the 15 genomes sequenced during the October 2000 Microbe Month effort; an additional ~30 genomes are anticipated to be processed as they become available in early 2002. Multiple gene callers (Generation, Glimmer and Critica) are used to construct a candidate gene model set. The conceptual translations of these gene models are used to generate similarity search results and protein family relationships; from these results a metabolic framework is constructed and functional roles are assigned. Simple repeats, complex repeats, tRNA genes and other structural RNA genes are also identified. Annotation summaries are made available through the JGI Microbial Sequencing web site; in addition, draft results are being integrated into the interactive display schemes of the Genome Channel/Catalog. Extensive use of high-performance computational tools has enabled rapid processing of genomes in batch. As of this writing, 22 genomes, comprising over 93 million bp of sequence, in ~4000 contigs have been processed to generate ~85,000 candidate peptide translations.

161. Annotation of Microbial Genomes Relevant to DOE's Carbon Management and Sequestration Program

F. Larimer[1], L. Hauser[1], M. Land[1], D. Hyatt[1], M. Shah[1], S. Stilwagen[2], P. Predki[2], D. Arp[3], A. Hooper[4], S. Chisholm[5], G. Rocap[6], B. Palenik[7], J. Waterbury[8], R. Atlas[9], J. Meeks[10], C. Harwood[11], R. Tabita[12], P. Chain[13], and J. Lamerdin[13]
[1]Oak Ridge National Laboratory, Oak Ridge, TN
[2]Joint Genome Institute, Production Sequencing Facility, Walnut Creek, CA
[3]Botany and Plant Pathology Department, Oregon State University, Corvallis, OR
[4]Department of Biochemistry, University of Minnesota, St. Paul, MN
[5]Departments of Civil and Environmental Engineering and Biology, Massachusetts Institute of Technology , Cambridge, MA
[6]School of Oceanography, University of Washington, Seattle, WA
[7]Scripts Institution of Oceanography, University of California San Diego, San Diego, CA
[8]Woods Hole Oceanographic Institution, Woods Hole MA
[9]Department of Biology, University of Louisville, Louisville, KY
[10]Section of Microbiology, University of California Davis, Davis, CA
[11]Department of Microbiology, University of Iowa, Iowa City, IA
[12]Department of Microbiology, Ohio State University, Columbus, OH
[13]Lawrence Livermore National Laboratory, Livermore, CA
fwl@ornl.gov

A diverse group of autotrophic microorganisms have been sequenced to further fundamental research into carbon management topics that would enable a reduction or slowed growth of the atmospheric concentration of carbon dioxide; potential routes include augmenting the natural carbon cycle by identifying ways to enhance carbon sequestration in the terrestrial biosphere through CO_2 removal from the atmosphere and storage in biomass and soils, and through evaluating the potential for increased carbon sequestration in the open oceans The aim of this research is to improve upon our rather rudimentary

understanding of how carbon is used and stored in the biosphere. By systematic analysis of each genome, we hope to identify specialized nutrient uptake systems, pathways that contribute to or regulate nitrogen utilization, carbon cycling and photosynthesis.

The target genomes comprise a diverse group of autotrophs that are significant in their respective ecosystems and contribute materially to cycling of atmospheric gases. Six genomes are being examined: three are marine cyanobacteria, *Prochlorococcus marinus* ecotypes MED4 and MIT 9313, and *Synechococcus sp.* WH8102; *Nostoc punctiforme*, a nitrogen-fixing fresh-water cyanobacterium; *Rhodopseudomonas palustris*, a metabolically versatile anoxygenic photobacterium; and *Nitrosomonas europaea*, an ammonia-oxidizing beta-proteobacterium. The three marine cyanobacteria and *R. palustris* are currently undergoing final annotation; *N. europaea* is at closure and *N. punctiforme* is in finishing.

These genomes comprise an extensive resource for comparative genomics: the cyanobacterial genomes, together with completed and ongoing cyanobacterial sequencing elsewhere, represent the first opportunity to deeply examine this form of photoautotrophy; *R. palustris* is extensively informed by the recently completed *Caulobacter crescentus*, *Sinorhizobium meliloti* and *Mesorhizobium loti* genomes, as well as the larger contiguous portions of the draft *Rhodobacter sphaeroides* genome, expanding the alpha-proteobacterial group.

(Research supported by the Office of Biological and Environmental Research, USDOE under contract number DE-AC05-00OR22725 with Oak Ridge National Laboratory, managed by UT-Battelle, LLC)

162. A Genome-Wide Search for Archaeal Promoter Elements

Enhu Li[1], **Aaron A. Best**[1], Gretchen M. Colon[2], Claudia I. Reich[1], and Gary J. Olsen[1]
[1]University of Illinois at Urbana-Champaign
[2]Gordon College, Barnesville, GA
abest@uiuc.edu

The archaeal basal transcription system is a simpler version of the eukaryotic system, having a single RNA polymerase (RNAP) and only two general basal transcription factors: TATA-binding protein (TBP) and transcription factor B (TFB). These factors bind specific promoter elements and recruit RNAP. Though consensus promoter elements and basal transcription factors have been identified in Archaea, it remains unclear how transcription is (*i*) initiated in the absence of canonical promoter elements or (*ii*) regulated.

We have adopted an iterative, genome-wide strategy to identify promoter elements in Archaea, using *Methanococcus jannaschii* as a model system. The strategy has isolated and identified 15 of the 23 predicted promoter regions for tRNA transcripts. Alignment of isolated tRNA promoters reveals near-consensus TATA-elements and TFB recognition elements (BREs) located within 100 nucleotides (nt) of the tRNA coding sequences. A third conserved element, possibly serving as archaeal Initiator, is located *ca.* 21 nt downstream of the TATA-element in most of the isolated tRNA promoters. Binding of TBP and TFB to the predicted promoter elements was confirmed by DNase I footprinting. The eight remaining tRNA promoters have been characterized by a targeted approach, and analyses reveal that three of these differ significantly from the consensus sequences. Electrophoretic mobility shift assays reveal that promoter elements deviating from the consensus are bound by TBP/TFB with lower affinities than promoters exhibiting the canonical pattern. In addition, a correlation between tRNA promoter strength and predicted codon usage was observed – promoters exhibiting a high degree of similarity to consensus sequences drive expression of tRNAs with correspondingly high codon usage. Experiments are currently underway to validate these observed trends. Generally, the search strategy selected strong tRNA promoters and some strong protein promoters. However, promoters with lower affinity for TBP and TFB have also been isolated, suggesting that this strategy will be useful in the identification of novel and/or sub-optimal promoter elements.

In addition to identification of archaeal promoters, we have addressed questions surrounding archaeal RNAP (*i*) structure and (*ii*) recruitment to promoters using *in vivo* and *in vitro* protein-protein interaction methods. (*i*) Archaeal RNAP subunit composition is similar to that seen in eukaryotic RNAPs. We have demonstrated that archaeal and eukaryal RNAPs adopt similar subunit architectures, extending

evidence of homology from the sequence level to quaternary structure interactions. (*ii*) Recruitment of archaeal RNAP to canonical promoters occurs through interactions between TFB and specific RNAP subunits. We have identified subunits of RNAP that contact TFB and propose a model for the DNA/TBP/TFB/RNAP transcription initiation complex.

163. New Markov Model Approaches to Deciphering Microbial Genome Function

John M. Logsdon, Jr.[1], Mark. A. Ragan[2], and Mark Borodovsky[3]
[1]Department of Biology, Emory University, Atlanta, GA 30322
[2]Division of Computational Biology and Bioinformatics, The Institute for Molecular Bioscience, The University of Queensland, Brisbane, Qld 4072 Australia
[3]School of Biology, Georgia Institute of Technology, Atlanta GA 30332
jlogsdon@biology.emory.edu

Upon development of efficient algorithms for gene finding in prokaryotic genomes it was observed that there are hundreds of genes that escape confident prediction unless special efforts are taken. The most interesting genes—often also the most difficult to predict—are those atypical genes whose DNA sequence features deviate strongly from the 'typical' ones. We have begun efforts to improve accuracy of predicting atypical genes in prokaryotes by Markov and Hidden Markov model based algorithms, such as GeneMark-Genesis and GeneMark.hmm. An important goal of this project will be the comparison of the predicted sets of atypical genes with sets indicated by alternative approaches (i.e. base composition bias methods). The algorithms will be extended to the analysis of genome draft sequences (nearly complete genomes) produced by high-throughput sequencing. Using atypical genes predicted by careful implementation of these new methods and estimated at hundreds per genome, a number of relevant biological questions will be addressed. Most importantly, we will use rigorous phylogenetic reconstruction methods to test the possibility that each atypical gene is a result of lateral (or horizontal) gene transfer (LGT), and, if so, from what lineage it was derived. We will, thus, identify the fraction of atypical genes that are bona fide LGTs, both across all genomes and within given genomes. With these analyses, we will be able to estimate the overall rates of LGT, particularly with respect to phylogenetic and/or ecological separation between donors and recipients. We will also determine, using comparative database methods, the putative functional roles of these atypical genes in order to better understand what types are most prone to be identified as atypical and, of those, which gene types are most likely to have been transferred between species. In particular, we will focus on those genes that have clear roles in the adaptation to or alterations of natural environments. Of all the genomes, the remaining, 'typical', set of genes (i.e. those which show little, if any, evidence of LGT) will be used to assess the validity of a phylogenetically stable 'core' of microbial genes. From these analyses, we plan to build a database of atypical genes and their inferred phylogenetic relationships for publicly available microbial genomes along with a web-based interface. This will allow its contents to be displayed and searched for specific genes, proteins and their phylogenetic relationships. These analyses and the resulting database will be a valuable resource for studies of microbial genome structural and functional evolution.

164. Genomic Plasticity in *Ralstonia eutropha* and *Ralstonia pickettii*: Evidence for Rapid Genomic Change and Adaptation

T. L. Marsh, S-H Kim, N. M. Isaacs, S. Eichorst, and K. Konstantinidis
Michigan State University, Center for Microbial Ecology and Department of Microbiology
MARSHT@msu.edu

We have begun an analysis on genomic plasticity in the genus *Ralstonia* using recently isolated strains of *R. eutropha* and *R. pickettii*. The former served as the ancestral strain in a long-term evolution experiment in which eighteen independent lineages were propagated for 1000 generations under two different environmental conditions. Dramatic

changes in both phenotype and genotype have been observed in the evolved lineages including large deletions to the genome. These deletions are being analyzed with an eye to identifying apparent preferred pathways in genomic degeneration. Regarding *R. pickettii*, 20 strains have been isolated from a 20 cm (depth) core of lake sediment contaminated with high concentrations of copper. All of these isolates are resistant to high levels of copper as well as several other heavy metals. The isolates display substantial differences in REP-PCR profiles, pulse field gel patterns, and plasmid content, suggesting significant genomic plasticity within a relatively small habitat volume. We report here on the sequence of a small plasmid detected in several *R. pickettii* isolates.

165. Lateral Gene Transfer and the History of Bacterial Genomes

Howard Ochman
Department of Ecology and Evolutionary Biology
University of Arizona Tucson, Arizona 85721
hochman@email.arizona.edu

Deriving meaningful information from complete genomes depends upon the comparisons between sequences. Therefore, an evolutionary framework is required for all stages of genome analysis and interpretation. We are using universally distributed molecular characters to resolve the relationships among bacterial lineages in an attempt to determine the evolutionary history and degree of gene transfer among bacteria. The objectives of the proposed research are to use existing published and newly determined nucleotide sequences of a large set of universally distributed genes among bacteria of differing degrees of genetic relatedness and to address the several questions relating to the role of gene transfer in shaping bacterial genomes. In addition to supplying the information about the extent of gene transfer, this research serves three additional functions: (1) The set of conserved genes adopted for these studies will provide a new framework for the identification and classification of bacteria spanning all levels of genetic divergence. (2) Sequence information from a common set of genes will allow, for the first time, direct comparisons of the rates and patterns of nucleotide evolution within and among bacteria. (3) Analysis of a defined set of genes yields a rapid measure of genome dynamics, makes use of the rapidly increasing number of incomplete, unassembled or unannotated bacterial genomes, and can be used to direct the focus of new sequencing endeavors.

166. Physiomics Array: A Platform for Genome Research and Cultivation of Difficult-to-Cultivate Microorganisms

Michel Marharbiz, William Holtz, Roger Howe, and **Jay D. Keasling**
Departments of Chemical Engineering and Electrical Engineering and Computer Science University of California Berkeley, CA 94720
keasling@socrates.berkeley.edu

The sequences of a number of microbial genomes have recently been completed or will be completed shortly. Many of these organisms contain novel genes, the function for which is not known. Further, many of these organisms have novel characteristics —such as the ability to transform abundant biopolymers into biofuels or the ability to remediate environmental contaminants—that make them important for DOE purposes.

The large cultivation parameter space that the researcher needs to explore to determine the function of novel genes, the optimal culture conditions for a desired bioconversion, or the most appropriate cultivation conditions for a previously unculturable microorganism is extensive. Given the large numbers of organisms that have been sequenced, unknown genes in each of those organisms, and previously unculturable organisms, a high-throughput cultivation device would allow one to explore cultivation parameter space quickly.

We are developing an integrated research program to develop a high-throughput physiomics array to assess the effects of changes in culture parameters or environmental contaminants on cell physiology and gene expression or to cultivate difficult-to-cultivate or previously unculturable microorganisms. The specific aims are as follows:

1. To develop a high-throughput physiomics array. This micro-system will be based on an array of 150-μl wells, each one of which

incorporates MEMS for the closed-loop control of cell culture parameters such as temperature, pH, and dissolved oxygen.
2. To test the effect of changes in culture conditions on growth rate, product formation, and contaminant degradation in *Deinococcus radiodurans* and *Shewanella putrefaciens*. We will measure cell density, substrate consumption, product formation, and contaminant degradation/accumulation.
3. We will analyze gene expression as a function of culture condition using gene chips

167. Optical Mapping: New Technologies and Applications

David C. Schwartz, Shiguo Zhou, Ana Garic-Stankovic, Alex Lim, Eileen Dimalanta, Arvind Ramanathan, Tian Wu, Ossmat Azzam, Casey Lamers, Brian Lepore, Aaron Anderson, Michael Bechner, Erika Kvikstad, Natalie Kaech, Andrew Kile, Jessica Severin, Rodney Runnheim, Danile Forrest, Christopher Churas, Galex Yen, Jonathan Day, Bud Mishra, and Thomas Anantharaman
University of Wisconsin-Madison, Department of Chemistry, Department of Genetics, UW Biotechnology Center
dcschwartz@facstaff.wisc.edu

Our laboratory has developed Optical Mapping, a system for the construction of ordered restriction maps from individual DNA molecules. Our work centers on the development of new systems for genome analysis, including Optical Mapping, which exploit novel macromolecular phenomena to answer important biological problems. These are built upon a complex mix of principles derived from multiple disciplines including chemistry, genetics, computer science, biochemistry, optics, surface science and micro/nanofabrication. Recently, "Shotgun" Optical Mapping was used to construct whole genome restriction maps of *Escherichia coli* O157:H7, *Deinococcus radiodurans*, and *Plasmodium falciparum* (the major causative agent of malarial disease) without the use of PCR, electrophoresis, or clones. Presently we are applying Shotgun Optical Mapping to the analysis of more complex genomes, including human and rice, as well as of numerous microorganisms, where our mapping efforts are offering new routes to understanding genome plasticity across closely related species. These efforts are also helping to facilitate the ongoing microbial sequencing projects at JGI, in terms of providing means for validation and aids for assembly. With the advent of a high-throughput Optical Mapping System, we are developing novel approaches for human association studies using a new class of genome markers that are designed to encompass SNPs (Single Nucleotide Polymorphisms), yet reveal genome variation on a scale not previously discerned for large populations. Current thinking in the field is centered on the use of a limited number of SNPs to leverage the apparent state of linkage disequilibrium, which is indicative of a young species; however, current approaches based on chips or mass spectrometry are pendant on huge numbers of oligonucleotides. This requirement limits analysis to a series of discrete loci and renders such approaches inadequate for the assessment of a broad spectrum of genome variation motifs. This limitation of current systems used for large-scale association studies may neglect discovery of important factors contributing to complex traits. In this regard, haplotyping is emerging as the means to perform detailed analysis of mutations and is expected to play a major role in the emerging field of pharmacogenomics. The Optical Mapping platform is uniquely suited for haplotyping since analysis of single molecules allows for the unambiguous phasing of genetic markers within populations of unrelated individuals.

168. Spectroscopic Studies of *Desulfovibrio desulfuricans* Cytochrome c3

William H. Woodruff[1], Judy D. Wall[2], Robert J. Donohoe[1], and Geoffrey B. West[1]
[1]Los Alamos National Laboratory
[2]University of Missouri, Columbia
woody@lanl.gov

Desulfovibrio desulfuricans is a sulfate-reducing bacterium that also is able to reduce a variety of metals including Cr(VI) and U(VI). Reduction in vitro with hydrogen as electron donor is dependent on the four-heme periplasmic cytochrome c3, a

broad-specificity redox protein. It is unknown whether cytochrome c3 acts as the proximate electron donor to the metal species, or whether it is simply an electron carrier in the respiratory redox network of this organism. We have undertaken characterization of cytochrome c3 by spectroscopic and other physical methods to establish the role of this protein in metal reduction. Resonance Raman results allow specific hemes and their redox states to be distinguished, and infrared results reveal the sidechain proton-transfer reactions that accompany the electron transfer steps. An allometric scaling model shows general correlations between genome length, average copy numbers of gene products, and bioenergetic capacity over a very large range of bacterial size.

169. Identification and Isolation of Active, Non-Cultured Bacteria for Genome Analysis

Cheryl R. Kuske, Susan M. Barns, Ellie Redfield, and Leslie E. Sommerville
Bioscience Division, M888, Los Alamos National Laboratory
kuske@lanl.gov

At least one third of the bacterial divisions identified to date have no cultured members. Non-cultured bacteria representing several bacterial divisions are widespread and potentially abundant in soils and other environments. For example, we have found that members of the Acidobacterium division are among the most abundant bacteria in some soils, yet we know almost nothing of their functions. The overall goals of our project are to determine the abundant and active members of the Acidobacterium division in pristine and contaminated soil and aquifer material using RT-PCR, 16S rRNA-targeted probes and in situ microscopy, and to collect cells of active, non-cultured groups by flow cytometry cell sorting. The pooled DNA of non-cultured bacteria isolated directly from the environment will be a valuable resource of genetic material for comparative analyses of conserved and novel gene families, and for targeted genome sequencing. Work in the last year has focused on technical advances in hybridization and flow cytometry separation of bacterial cells from natural environments. To apply these techniques to analysis of cells from soil and to enrich for bacterial groups of interest, we are comparing the bacterial diversity found in pools of bacterial cells fractionated from soil with that of the parent environment. We have also begun work on RT-PCR methods for analysis of active Acidobacterium division members from contaminated and pristine soils.

170. Assembly of Microbial Sub-Genomes from Beneath a Leaking High-Level Radioactive Waste Tank

Fred Brockman, Margie Romine, Greg Newton, Amber Alford, Shu-mei Li, Jim Fredrickson, Kristen Kadner, Paul Richardson, and Paul Predki
Pacific Northwest National Laboratory and DOE Joint Genome Institute
fred.brockman@pnl.gov

Our goal is to demonstrate the ability to obtain 1 to 2 Mbp of genetically linked sequence (a sub-genome) from microorganisms that can not be grown in pure culture by direct cloning of DNA from environmental enrichments and high throughput sequencing of BAC ends. Simulations indicate that paired-end sequencing of approximately 5000 BACs from a well-represented library where 5 to 10% of the bacterial community is composed of an "archetypal species" (a single species or a closely related group of species) could produce a contig of 1-2 Mbp before chromosome walking fails. Subsurface vadose zone (aerobic) sediment samples—representing the most radioactive sediment samples ever taken at the DOE Hanford Site in Washington state—are the focus for this demonstration. Samples contained up to 50 microCuries of Cesium-137 per gram sediment, other radionuclides at nano- and picoCurie levels, and pH's to 9.8. Samples in which no microorganisms could be grown on solid media but which produced growth in pH 10 and/or 50 degree C liquid media enrichments were selected for study. In the first several months of the project, the microbial community in these enrichments and subsequent transfers have been screened for species that comprise >5% of the community and for bacterial divisions with few or zero cultured representatives, as a basis for determining appropriate community(ies) for BAC library construction.

171. The Marine Environment from a Cyanobacterial Perspective

Brian Palenik[1], Ian Paulsen[2], Bianca Brahamsha[1], Rebecca Langlois[1], and John Waterbury[3]
[1]Scripps Institution of Oceanography, University of California, San Diego
[2]The Institute for Genomic Research
[3]WHOI
bpalenik@ucsd.edu

The genome sequence of the marine cyanobacterium *Synechococcus* strain WH8102 is nearly completed. This microorganism was chosen because cyanobacteria similar to WH8102 are ubiquitous and significant primary producers in oligotrophic marine environments. In addition this strain possesses a unique type of prokaryotic motility and is amenable to genetic manipulation. The genome is estimated to be 2.7 Mb with approximately 2390 ORFS. The transporter complement of *Synechococcus* WH8102 was analyzed by screening its genome against a database of known and putative transporters by BLAST and HMM-based analyses. Approximately eighty transport systems were identified comprising 130+ genes. Comparison with the transporter complement of other complete genomes indicated that WH8102 has an emphasis on transport of inorganic anions, in particular with multiple transporters for nitrate, sulfate and chloride. In terms of organic nutrients it is predicted to transport a variety of amino acids and a limited number of sugars. The transporters and other activities of the cell are coordinated by a surprisingly small number of two component regulatory systems compared to the freshwater cyanobacterium *Synechocystis* PCC6803. Ultimately the WH8102 genome will provide us with a better understanding of how cyanobacteria perceive and respond to the marine environment.

172. Metagenomic Analysis of Uncultured Cytophaga and Beta-1,4 Glycanases in Marine Consortia

David L. Kirchman and Matthew T. Cottrell
College of Marine Studies, University of Delaware
kirchman@udel.edu

Culture-independent studies have shown that microbial consortia in natural environments are incredibly diverse and are dominated by bacteria and archaea substantially different from microbes maintained in pure laboratory cultures. Recent studies indicate that previous culture-independent studies using PCR-based methods have largely overlook an important group of uncultured bacteria, the *Cytophagales*. These bacteria appear to be abundant in the oceans and probably other oxic environments. We hypothesize that the key to understanding consortia and their function in organic matter mineralization in oxic environments is to focus on uncultured *Cytophagales* and their genes encoding endoglycanases. This poster will summarize our progress in understanding uncultured *Cytophagales* and our plans for our new DOE-supported metagenomic project. We have been using an approach that combines microautoradiography with fluorescence in situ hybridization (Micro-FISH) to examine which bacterial groups are responsible for using naturally-occurring organic material. As we had hypothesized based on the work with cultured representatives, uncultured *Cytophagales* appear to dominate use of protein and chitin in the Delaware Bay and coastal waters. Perhaps as a result of protein and chitin inputs, uncultured *Cytophagales* are abundant through the Delaware estuary. For our new project, we intend to construct two BAC libraries with DNA directly (no PCR) from uncultured microbial consortia found in a coastal marine environment. Microbes on macroscopic aggregates will be one target for our clone libraries. These organic aggregates, which are important in carbon transport and storage in the oceans, harbor dense assemblages of *Cytophagales*. The libraries will be screened for 16S rRNA genes, for cellulase and chitinase-active clones, and for clones bearing genes of these enzymes. The DNA probes for screening the libraries will be constructed from the sequence data now emerging from the *C.*

hutchinsonii project which has already found about 15 presumed endoglucanases (mainly cellulases). The proposed work should reveal much about a neglected microbial group that appears to dominate microbial assemblages in oxic environments. Ultimately, the data will be used to improve models of carbon cycles and storage in the oceans and other environments where *Cytophagales* are abundant and ecologically important.

173. Rational Design and Application of DNA Signatures

P. Scott White, John Nolan, Rich Okinaka, Paul Jackson, and Paul Keim
Bioscience Division, Los Alamos National Laboratory and Department of Microbiology, Northern Arizona University
scott_white@lanl.gov

With the rapid accumulation of direct sequence data for a variety of pathogenic organisms, the development and application of pathogen "signatures" is undergoing a paradigm shift from empirical development of signatures using detection platform-specific methods, to rationally designed signatures that can be assessed in a platform-independent manner. Thus, DNA sequence is the "signature", and the signatures have precise phylogenetic and functional significance. We are using phylogenetic and functional analysis, combined with a rapid method for direct sequence analysis using microsphere arrays and flow cytometry, to exploit the information contained in DNA sequence from multiple genetic loci. Such Multi-Locus Sequence Typing (or MLST) has the potential to revolutionize DNA-based analysis in applications ranging from biological point detection to water and food safety. Currently we are focusing on DNA sequence analysis tools (bioinformatics), the design of DNA primers and probes (reagent development), and protocol development for both laboratory and field applications.

174. Pathogen Detection: Successes and Limitations of TaqMan® PCR and Limitations of TaqMan® PCR

Shea N. Gardner, Thomas A. Kuczmarski, Elizabeth A. Vitalis, and Tom Slezak
Lawrence Livermore National Laboratory
gardner26@llnl.gov, tomk@llnl.gov

Recent events illustrate the imperative to rapidly and accurately detect and identify pathogens during disease outbreaks, whether they are natural or engineered. Detection techniques must be both species-wide (capable of detecting all known strains of a given species) and species specific. Fluorogenic probe-based PCR assays (TaqMan®; Perkin Elmer Corp./Applied Biosystems, Foster City, Calif.) may be a sensitive, fast method to identify species in which the genome is conserved among strains, for example, in West Nile/Kunjin virus. For species such as Venezuelan Equine Encephalitis and HIV, however, the strains are highly divergent. We use computational methods to show that 6-10 TaqMan® primer/probe sequences, or signatures, are needed to ensure that all strains will be detected, an unfeasible number considering the cost of TaqMan® probes. We compare TaqMan® with the alternate nucleic acid based detection techniques of microarray, chip and bead technologies in terms of sensitivity, speed, and cost.

175. Sequencing and Analysis of the Genome of *Carboxydothermus hydrogenoformans*, a CO-Utilizing, Hydrogen Producing Thermophile

J. A. Eisen[1], F. T. Robb[2], J. Gonzalez[2], T. Sokolova[3], L. J. Tallon[1], K. Jones[1], A. S. Durkin[2], and C. M. Fraser[1]
[1]The Institute for Genomic Research, Rockville, MD
[2]University of Maryland Biotechnology Institute, Baltimore, MD
[3]Russian Academy of Sciences, Moscow, Russia
jeisen@tigr.org

Carboxydothermus hydrogenoformans is an extreme thermophilic bacterium, growing on CO as the only

carbon and energy source under strictly anaerobic conditions. Here we present an update on the progress of sequencing and analyzing the genome of this species. Preliminary analysis reveals that this species is clearly a low-GC gram-Positive bacteria in most aspects of its core biology. However, this species encodes many genes, in particular those likely involved in energy metabolism, that are more commonly found in distantly related thermophilic or methylotrophic bacteria and Archaea. Analysis of various features of the genome will be presented.

Ethical, Legal, and Social Issues

176. Intellectual Property Rights Issue Concerning the Human Genome: A Test of Anticommons Theory and Implications for Public Policy

David J. Bjornstad[1] and Steven Stewart[2]
[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee
[2]Department of Economics, University of Tennessee, Knoxville
dub@ornl.gov

Concerns that patenting policies for genetic research may lead to inefficient use of genetic research has come about through a number of circumstances. These circumstances include patent policy, technological changes in gene sequencing, rapid entry of the private sector into gene sequencing research, and the changing relationship between government, the traditional non-profit research sector, and the profit-motivated private sector. The larger project from which the current poster is drawn will deal with all of these subjects. The information presented in this poster reports on one phase of this work, an "experimental" test of the theory of the "anticommons." Anticommons theory postulates that an overassignment of property rights will lead to an underutilization of a resource, an analogy to the tragedy of the commons in which underassignment of property rights leads overutilization of a common pool resource. Applied to the human genome topic this theory suggests that the patenting of gene fragments will lead to a circumstance in which the information developed through gene research will be underutilized. This poster describes empirical research on this topic and its implications for patent policy.

177. Regulation of Biobanks: Banking Without Checks or Insured Deposits?

Mark A. Rothstein[1], Bartha M. Knoppers[2], **Mary R. Anderlik**[1], Genevieve Cardinal[2], and Mylene Deschenes[2]
[1]Institute for Bioethics, Health Policy and Law, University of Louisville
[2]Centre de Recherche en Droit Public, University of Montreal
mrande01@gwise.louisville.edu

Between the high-tech, abstract science of genomics and the practical, gene-based medicine of the future lies an emerging world of biomedical research involving pharmacogenomics, genetic epidemiology, and numerous other scientific specialties. Unlike genomics, in which a few dozen DNA samples supported a world-wide research enterprise, the next stage of research requires thousands or even millions of biological specimens. There is now a rush to compile huge repositories of biological specimens linked to phenotypic data. The significant questions of law, ethics, and policy raised by biological sample collections are complicated by the emergence of commercial biobrokers, the population-based and international scale of the new specimen banks, and the use of the Internet in recruiting donors. The research team is developing a conceptual framework for understanding and regulating traditional and emerging forms of biobanking.

178. Healthy, Working Individuals' Perspectives on Ethical, Legal and Social Issues Involved in Complex Genetic Disorders

Teddy D. Warner, Melinda Rogers, Julienne Smrcka, Nashe Garcia, Kate Green Hammond, Cynthia Geppert, and Laura W. Roberts
Department of Psychiatry, University of New Mexico School of Medicine
tdwarner@salud.unm.edu

This pilot project examines healthy working peoples' perceptions of ethical, legal and social issues concerning complex genetic disorders (e.g., alcoholism, cancer, depression, diabetes). Over the past year, our multidisciplinary team has constructed and pilot-tested (N = 60) a new comprehensive survey instrument (1.5 hours duration; 65 pages; 490 items) to assess a wide range of attitudes and perceptions regarding genetically-related disorders, genetic testing, research and clinical use of genetic information (e.g., confidentiality, disclosure of information, informed consent), concerns about risks of exposure to various agents, influence of various factors in specific work or health-related scenarios, use of genetic information by various organizations, experience with genetic disorders or genetic testing, and other related issues. The quantitative measures from the survey allow comparison of attitudes about personal genetic information gathered for clinical versus research purposes. A 20-minute interview precedes the written survey and gathers qualitative responses to 17 questions (interviews were videotaped with consent to record responses and to enable use in scientific presentations). For this presentation we will characterize the responses of 60 participants, half from Sandia National Laboratories (of the Department of Energy) and half from the University of New Mexico Health Sciences Center (a comprehensive medical center). Preliminary information from this sample will be used to support our proposal next month to DOE to execute a large scale (about N = 900), hypothesis-driven project.

179. GeneTests•GeneClinics: A Primer for Non-Geneticists

Roberta A. Pagon
Department of Pediatrics, University of Washington School of Medicine, Seattle, WA
bpagon@u.washington.edu

The recently merged GeneTests•GeneClinics (www.genetests.org or www.geneclinics.org) is a publicly funded medical genetics information resource developed for physicians, other healthcare providers, and researchers, available at no cost to all interested persons. Although GeneTests•GeneClinics was initially developed as a resource for genetics professionals, the vast majority of the 30,000 registered users who access GeneTests•GeneClinics each day are not formally trained in genetics and include numerous healthcare providers, most of whom have had a limited exposure to medical genetics. The purpose of this project is to develop short, concise definitions, illustrations, and vignettes explaining a term or concept that can be viewed in context by GeneTests•GeneClinics users without leaving the text they are reading. The GeneTests•GeneClinics Web site currently has a glossary of over 220 words, which is not displayed in context and can only be accessed by clicking on "Educational Materials" on the Web site navigation bar. As our grant funding had not yet started at the time this abstract was submitted (11/15/01), project staff has not yet been hired and the project is still in preliminary planning stages. In anticipation of a late November - early December start date, we will develop prototype definitions, illustrations, and vignettes to be evaluated by non-geneticists before this meeting in order to obtain feedback from the target audience early in the project. Examples developed to date will be displayed either as a poster format or live demonstration.

180. *Science and its Appeals*

Noel Schwerin
Backbone Media, 58 Harper Street San Francisco, CA 94131
schwerin@backbonemedia.org

Science and its Appeals (previously called *Truth and Justice*) is a one-hour documentary produced by award-winning Backbone Media for national broadcast on PBS. The first of its kind, *Science and its Appeals* will explore how ordinary individuals poised at the intersection of two of our most powerful institutions - science and the courts - are struggling to make sense of the fundamental questions raised by new genetic, reproductive and life-creating technologies: What is a human being? Who is a parent? Who owns your body? Who has a right to know its secrets?

Science and its Appeals will profile individuals – lay people and lawyers, judges and scientists – as they grapple with questions of new technology and the law in a handful of current – *rather than hypothetical* – legal cases. It will demonstrate how new technologies create unexpected, unprecedented dilemmas that force us to rethink fundamental ethical, social and legal principles. Through these

central "characters," *Science and its Appeals* will help articulate more fully and precisely the real (and surprisingly fundamental) questions underlying our visceral reactions to new genetic, reproductive and life-creating technologies.

When people react strongly to the possibility of patenting a part-human animal, creating a child with possibly eight parents, or being tested secretly for genetic disease, they are not always reacting out of fear or ignorance, but rather, they are responding in part to the profound nature of the technological challenge. Who are we as a species and how is that changing? Do advances in biotechnology confuse or clarify what it means to be human, to be a parent, to inhabit a body? How do we balance scientific progress with other values? Are there things that are scientifically possible, medically justifiable and legally allowed, but that might also prove socially objectionable? If so, how will we renegotiate long-held paradigms and belief systems?

Despite the profound impact new life technologies are having on the way we feel in the world, and the role of the courts in defining and mediating that impact, no PBS – or network – series has ever explored the topic beyond a discussion of hypotheticals or the predictable recitation of what might go wrong with new technology. Yet when we worry about transgressing natural boundaries, exercising too much power over the natural world or not being able to protect ourselves or our loved ones from harm, we are engaging not only in philosophical thinking, we are revisiting the very foundations of our laws and social contracts. The compelling and *real* stories of *Science and its Appeals* will better articulate this direct relationship between what we feel, the fundamental, philosophical questions embedded in those feelings, and the legal institutions and social relations we have built to honor them.

181. Convergence

Cynthia Needham and Kenneth McPherson
ICAN Productions, Ltd., Stowe, VT
cynthia@smartscience.org

The fields of molecular biology, computational and materials engineering, chemistry, and physics are converging along a common path called nanotechnology – a path yielding unprecedented understanding and control over the fundamental building blocks of all physical things, both animate and inanimate.

While the science is powerful and intriguing, it also raises compelling social, legal and ethical issues. ICAN Productions, in association with Oregon Public Broadcasting, Palfreman Film Group (PFG), and the National Association of Biology Teachers, is developing a multi-component project focusing on the field. Through the support of the Department of Energy's Office of Biological and Environmental Research, ICAN Productions proposes to do three things to further the project's development. First, we will conduct interactive dialogs with research scientists in the respective nanotechnology disciplines to choose the specific research efforts that will delineate the most important ELSI issues. Second, we will develop similar discourses with senior scientists, ethicists, and legal scholars to explore their views on whether our current frameworks for ethics and justice support decision making within this new scientific enterprise. Third, we will work with all participants to identify stories that will provide a compelling public viewing experience while best illustrating both the science and the ethical, legal, and social issues that we have identified. These core stories will form the structure upon which each component of the project is based.

Work to date has allowed ICAN and PFG to identify several key scientific themes that will likely drive the stories. Some, if not all, of the themes extend the extensive genomic research and its products into a new dimension. Health scientists envision 'nachines' that will patrol our bodies, searching out and destroying invading microbes or pre-cancerous cells, new abilities to repair damaged genes or cells molecule by molecule, and ultimately the ability to enhance or improve upon an individual's natural assets. Materials scientists and chemists promise smart materials that will change as their environment changes and materials that will be hundreds of times stronger and lighter than any that now exist. Engineers promise faster and faster computer processors with vastly increased information storage to analyze, design, and control every aspect of the world at the nanoscale, breaking through the restrictions imposed by silicon. And finally,

biologists are charging toward the ultimate – the creation of life itself. Each of these themes offers opportunities to probe our present constructs of ethics and social justice.

182. Delivering the Human Genome to the Public

Sara L. Tobin[1] and Ann Boughton[2]
[1]Program for Genomics, Ethics, and Society, Stanford Center for Biomedical Ethics
[2]Twisted Ladder Media
tobinsl@stanford.edu

Progressive identification of new genes and implications for medical treatment of genetic diseases appear almost daily in public media reports. However, most of the public lacks the genetic sophistication to appreciate these advances or to anticipate the impact that the Human Genome Program will have on their lives. This lack of education may delay their acceptance of new medical options, limit their ability to communicate with medical providers, and lead to uninformed fears of products involving molecular technologies. This project is designed to fill two important functions: first, to provide user-friendly education for the public about genomics and molecular technologies, including the impact, implications, and potential of this field for the treatment of human disease; second, to contribute to the creation of an informed electorate with an appreciation for technological research and its benefits.

On the basis of previous support from the DOE Human Genome Program, we have produced two flexible, user-friendly, interactive multimedia CD-ROMs about the applications of molecular medical genetics. Physicians who use "The New Genetics: Courseware for Physicians" can apply for continuing medical education credits through the Stanford University School of Medicine. The second version, "The New Genetics: Medicine and the Human Genome," is being used by members of diverse audiences, including a graduate class in the genetics of speech and hearing disorders, internal medicine residents at the Mayo Clinic, and training programs for technical personnel for the biotechnology industry. This current project builds on the content of these two multimedia CDs. However, the content is being adapted for delivery on the Internet and also modified so that members of the general public will find it easy to understand. The emerging website is designed to provide education in four areas: (1) Genetics, including DNA as a molecular blueprint 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 issues such as privacy and impact on the family. The website that results from this project will be made freely available to the general public and is designed to provide a powerful tool for education about the potential of the Human Genome Program to benefit human health, technological careers, and economic growth.

183. Major Psychiatric Diseases: A Model for Teaching Genetics Professionals about Complex Disorders

Joseph D. McInerney and Holly L. Peay
National Coalition for Health Professional Education in Genetics
jdmcinerney@nchpeg.org

The National Coalition for Health Professional Education in Genetics (NCHPEG) is developing an interactive, educational CD-ROM on psychiatric genetics that will be distributed free-of-charge to all active members of the National Society of Genetic Counselors (NSGC), a member organization of NCHPEG, and to all training programs for genetic counselors in the US and abroad. Additional copies will be available by request to all NCHPEG member organizations and to individuals who do genetic counseling.

The program addresses the following topics:

- descriptions, prevalence, and natural histories of schizophrenia (SZ), major depressive disorder (MDD), and bipolar disorder (BPD);
- current approaches to and limitations of diagnosis and treatment, and the implications for genetic counseling;
- current understanding of disorder etiology;

- the status of research into genetic contributions to SZ, MDD, and BPD, including a review of research methodologies;
- current approaches to genetic counseling for SZ, MDD, and BPD including psychosocial issues in genetic counseling for psychiatric disorders and family history evaluation;
- the implications for psychiatry and genetic counseling of potential gene discovery and genetic testing;
- the importance of collaborative relationships between mental-health professionals and genetics professionals in providing psychiatric genetics services;
- ethical, legal, and social issues that arise from continued research into the genetic basis of psychiatric disorders; and
- the ways in which these adult psychiatric disorders demonstrate the growing importance of common, complex diseases in genetic medicine.

Joseph D. McInerney is principal investigator for the project. Holly Landrum Peay, a board-certified genetic counselor and member of NSGC, is the project director.

The development process includes continued input from experts in the field of psychiatry, genetics, ethics, and genetic counseling. In November 2000 the project advisory committee determined the initial framework for the CD-ROM content, and in February of 2001 the writing committee wrote the bulk of the CD-ROM materials. During this time we also surveyed NSGC members to determine what educational materials they felt would help them counsel for psychiatric indications, and we have made every effort to include those materials.

During the spring of 2001 we expanded the content developed during the first writing meeting and our CD-ROM development consultant placed the materials on the CD-ROM. This first test (or alpha) version of the CD-ROM was evaluated by our target audiences. We have recently completed testing of the program with approximately 75 practicing counselors in various parts of the country and with students in training programs in the U.S. and abroad. In addition, the program was tested by members of the International Society of Nurses in Genetics (ISONG), a NCHPEG member organization. We compiled and analyzed information from the field test, and presented the information to the advisory committee during the second meeting in November 2001. Based on these data, the advisory committee suggested changes to the CD's content and structure. These changes will be implemented at the January 2002 meeting of the writing committee.

184. *THE AGE OF GENES - The Science of Your Life in the New Genomic Era*: A Television Series and Journalism Education Project

Peter Baker[1] and Barbara Wold[2]
[1]SeeingScience Media Group
[2]California Institute for Technology
peterrbaker@earthnet.net

Completion of the central endeavor of the Human Genome Project (HGP)—learning the DNA sequence of the entire human genome—heralds the beginning of a new phase of science that promises to touch the lives of all Americans in a way not equaled since the splitting of the atom. However, despite the HGP's overwhelming implications for divers aspects of life, the public finds itself perplexed by the rapid pace and complexity of genomic research and, quite understandably, confused about its ramifications. The vast majority of working journalists, whose job it is to translate this research into language comprehensible to a lay audience, are themselves woefully unprepared for the task. The "Age of Genes" project, based at the California Institute of Technology, and scientifically led by a team of world class genome scientists and biologists, takes a two-pronged approach to this problem. The project consists of: (1) a high-quality, four-part television series for national prime-time broadcast developed by the award-winning production company of Baker & Simon Associates (SeeingScience Media Group), aimed directly at educating a vast public audience on the scientific and social dimensions of cutting edge genomic research; and (2) a series of Seminars and Institutes, and the Web site FACSNET, offered by the Foundation for American Communications (FACS), to help journalists translate cutting edge genomics research into language accessible to the

average person through the print and electronic media which they serve. The television series, augmented and supported by the program of journalist education, has the potential to educate millions of Americans about the science and the challenges of the new genomic era.

185. Information Conferences on the Human Genome Project

Kathryn T. Malvern and Issie L. Jenkins
Zeta Phi Beta Sorority, Inc.
Drktmalvern@aol.com

Since 1997 the Zeta Phi Beta Sorority National Educational Foundation has been involved in bringing to minority communities information on the Human Genome Project and the resulting ethical, legal, and social issues. Through major conferences, and smaller workshops and seminars, the Foundation has been able to reach minority communities throughout the country, to provide information on genetic research developments, to obtain community input and recommendations regarding community concerns, and to help to encourage minority students in careers in science and biotechnology.

The Foundation has presented major conferences in New Orleans, Philadelphia, Atlanta, and Washington, D. C. In addition, it has collaborated in the presentation of smaller workshops and seminars. Participants at the most recent major conferences in Philadelphia in July 2000, in Atlanta in July 2001, and in Washington, D. C. in November, 2001, continue to stress the community needs for information on genetic developments, and for community input.

The Foundation's presentation/poster will highlight the Foundation's experiences in getting information out to minority communities, the optimism and concerns that minority communities have, as expressed by conference participants via recommendations, and the various local community outreach efforts that have resulted from the conferences.

The conferences have included participants representing the African-American community, the Hispanic community, the Asian American community, and the Native American community.

One of the Foundation's Goals is to encourage participants to take the information back and present it to their groups and organizations, helping to reach all segments of the community.

Pursuant to community requests resulting from the information conferences, the Foundation has assisted groups to organized other workshops and presentations on the Human Genome Project in their local communities. Ten mini-grants have been provided to help local groups with workshops and presentations. During the past year, the Foundation has collaborated with the Pennsylvania Legislative Black Caucus in organizing a Summit in Pennsylvania on the Human Genome Project, where conference participants made twelve recommendations to the legislators; organized and presented an ELSI workshop at the national IMAGE, Inc. convention, an organization representing Hispanic Americans; provided information at educational forums of community groups and at science and math fairs for students; as well as given university lecture on the implications of the Human Genome Project.

186. Assessing Models of "Public Understanding" in ELSI Outreach Programs

Bruce Lewenstein
Cornell University
b.lewenstein@cornell.edu

For more than a decade, outreach projects funded under the "Ethical, Legal, and Social Issues" (ELSI) rubric of the Human Genome Project have provided a base for public awareness, learning, and discussion of emerging genome science. Over that same period, new concepts of "public understanding" have emerged, moving from a "deficit" or linear dissemination model of popularization to a contextual model stressing lay knowledge, public engagement, and public participation in science. This project uses the base of ELSI projects to explore the ways that information about a new and emerging area of science that is intertwined with public issues has been used in public settings (including educational ones) to affect public understanding of science. By combining retrospective evaluation with new conceptual models of public understanding, this project moves beyond

evaluation of individual projects to examine the overall impact of ELSI projects on public understanding of genome science. This project combines detailed case histories of ELSI outreach projects with a comprehensive review of new concepts in public understanding.

187. Initiatives in Equity

Maria Elena Zavala[1,2], Lin Hundt[2], Jerry Beat[2], and **Marina Bobadilla**[2]
[1]California State University, Northridge
[2]Society for the Advancement of Chicanos and Native Americans in Science, Santa Cruz, CA
lin@sacnas.org

The Society for the Advancement of Chicanos and Native Americans in Science (SACNAS) seeks to increase the participation of minorities in the scientific endeavor. To address the needs of minorities in the sciences, SACNAS provides creative approaches to improving science education and equalizing opportunities in the national scientific work force. SACNAS delivers quality mentoring and professional development through its national conference, summer research programs, and publications. The Society's strength lies in the active involvement of its members, a dedicated board of directors, and a strong multilevel network between federal agencies, professional scientific societies, universities, and the private sector. For 27 years, SACNAS and its partners have succeeded in accomplishing the Society's mission of encouraging Chicano/Latino and Native American students to pursue graduate education and obtain the advanced degrees necessary for research careers and science teaching professions at the highest levels. The primary means of fulfilling the SACNAS mission is its annual conference. The SACNAS National Conference provides a forum for students, faculty and professionals in science and education to share research, and address the unique accomplishments and challenges of minorities in science. In conjunction with efforts from science government agencies, universities, professional societies, and private industry, the national conference that creates a forum to: 1) build networks on a national level among conference participants; 2) mentor undergraduate and graduate students; 3) expose participants to cutting-edge scientific research, current trends and issues; 4) inform participants about summer programs, internships, higher education, and employment; 5) address the professional demands and concerns of minority students and faculty; and, 6) empower K-12 educators to engage minority students in inquiry-based math and science. The SACNAS National Conference is the largest single outreach effort of the year and is the main vehicle by which the society fulfills its vision.

188. Creating and Distributing *Your World* Materials about Microbial Genomics

Jeff Alan Davidson[1], Cathryn Delude[2], and Ken Mirvis[2]
[1]Biotechnology Institute and BioSciEd
[2]The Writing Company
JeffDavidson01@cs.com

With the Department of Energy's support the Biotechnology Institute (BI) is developing a set of materials on Microbial Genomics for distribution in April of 2002 that will feature:

- a special *Your World* issue,
- an accompanying poster,
- a *Teacher's and Students Guide*, and
- an accompanying demonstration experiment.

Your World, is a magazine of biotechnology applications designed for teachers to use with 7th to 12th grade students. *Your World* is an eleven-year old program and has met with exceptionally positive reaction from students and teachers and 19 issues of the publication have been published on a wide range of biotechnology topics.

The *Your World* Microbial Genomics issue will be a colorful sixteen-page material designed for students use and will cover both science concepts and the application of those science concepts in the classroom. The issue will feature an introduction, a science overview article, several applications articles, an experiment that explores the central theme of the issue, and a profile of a scientist active in the field. Other elements include "Think About This" questions, Career Connections that highlight

career possibilities, and articles or sidebars that set up discussions among students about ethical issues.

The poster will be designed for both classroom display and to serve as a means of making science teachers aware of the availability of the issue on microbial genomics. The *Teacher's Guide* will provide additional background material, teacher resources, science standard connection information, and information on conducting the accompanying experiment. The *Student Guide* will provide information for the students on the conducting of the experiment including appropriate preparation information, data forms and data analysis recommendations. The accompanying experiment will be designed to allow students to experiment in a meaningful way to explore microbial genomics.

This project addresses middle school and high school biology teachers and the students they teach, and will provide approximately 5,000 to 8,000 sponsored or subscribing teachers with a complete set of classroom teaching materials on microbial genomics. Each subscribing teacher will receive 30 copies of *Your World* and copies of the supporting materials, and thus approximately 150,000 to 240,000 copies of the *Your World* issue will be distributed. Each set can be used with multiple classes and can be inventoried for reuse. This project is designed to help teachers introduce microbial genomics to approximately 800,000 students.

The Biotechnology Institute (BI) is a national non-profit entity, based in Washington, D.C., dedicated to education and research about the present and future impact of biotechnology. Its mission is to engage, excite and educate as many people as possible, particularly young people, about biotechnology and its immense potential for solving human health and environmental problems.

189. Modeling The Science and Technology Reference Court (STREC)

Franklin M. Zweig
Einstein Institute for Science, Health and the Courts, 2 Wisconsin Circle, Suite 700, Chevy Chase, Maryland 20815

In the course of genomics education provision to 3,100 judges supported by DOE's Ethical, Legal and Social Issues Program during the past four years, discussions among judges and scientists have concluded that a new institution is needed to bridge the gap between science and the law as the world adjusts to biotechnologies built upon the Human Genome Map and Sequence. Accordingly, *EINSHAC* initiates an evaluated model-building effort to design, initiate, test and assess a science and technology reference court (STREC) for high-profile biology and life technology disputes arising from human, mammalian, microbial, environmental and agricultural genomics and related ethical, legal and social issues.

The Science and Technology Reference Court will be simulated in Ottawa, Canada in June 2002 and again in Melbourne Australia, November 2003. These simulations will test the concept with real and hypothetical cases. Partners for these simulations are the Supreme Court of Canada; the Federal Court of Australia; the American Society for Human Genetics; The Society for Neuroscience; and a host of science centers in the United States and across the globe.

A reference court is limited to advisory verdicts, rulings and decisions. It does not seek enforcement of its conclusions. It enables official adjudication entities to review the best, considered judgments of neutral, independent scientific and technological expertise channeled through time-honored methods of judicial review. Legislative and administrative entities may also refer cases or questions.

The STREC promotes active scientific and jurisprudential collaboration at the highest levels, a primary aid to more effective, better-understood, more robustly supported, official systems of criminal and civil justice. At the same time it systematically reduces the disputed issues – safety of genetically modified organisms in a new technology, for

example – to assessed parameters of scientific certainty and uncertainty. It thereby may lower the rancor often incident to adversarial proceedings and promote areas of agreement in service to dispute resolution.

Using a variety of techniques, including summary jury trials and made-for-decision risk assessments, the STREC will adjudicate six cases types: (1) safety of genetically modified foods; (2) bioremediation projects using genetically modified bacteria; (3) human experimentation in biological agent attack preparation; (4) liability of professional societies for failure of national legislatures to protect human rights in genetics programs; (5) human cloning; (6) gene therapy for modifying human brain function.

The model project will recruit, train, orient and activate an advisory high court and constituent three judge panels to articulate scientifically respectable, legally supportable advisory opinions. In reaching its opinions, decisions, judgments, verdicts and rulings, both the nine member STREC and the three judge panels will have constant access to science advice. Six, world-recognized, jurists and three global science leaders will comprise the STREC. Senior and retired judges from different nations will comprise the STREC's reference panels, three-judge forums that examine disputes, issues opinions, and undertake the lion's share of case management. Reference panels will be coordinated by a STREC administrative officer. The American Bar Association's Model Canons of Judicial Ethics will otherwise provide guidance for all STREC members and reference panel jurists. A prominent scientist will chair the STREC's Science and Technology Commission.

STREC's roots, while respectful and incorporative of science, are historically judicial. They lie in the Congressional Reference jurisdiction of the U. S. Court of Federal Claims, a national court acting to advise Congress since 1856 according to federal statute, a policy court advisory pedigree discussed in the proposal narrative. STREC's legitimacy also receives support from the advisory powers of the Supreme Courts of eleven States of the United States, detailed in the proposal narrative. These developments indicate an attitude change within the Judicial Branch, recently noted by U.S. Supreme Court Justice Stephen Breyer:

"In this age of science, we must build legal foundations that are sound in science as well as in law. Scientists have offered their help. We in the legal community should accept that offer." – Hon. S. Breyer *Introduction to Reference Manual on Scientific Evidence*, second edition, Washington, DC: Federal Judicial Center, Administrative Office of U. S. Courts.

190. Ethical and Legal Issues Arising from Complex Genetic Disorders: The Law's Assessment of Probabilities

Lori Andrews, Laurie Rosenow, and Valerie Gutmann
Chicago-Kent College of Law
landrews@kentlaw.edu

As part of a larger project on the ethics of genetic testing for complex, common diseases, this preliminary study analyzed the way in which courts in cases involving negligence law and discrimination law have addressed genetic testing and genetic disease. The study attempted to predict whether complex genetic diseases will be handled differently than single gene disorders.

Under the American with Disabilities Act (ADA), people who have a record of, are regarded as having, or do have a disability are protected from discrimination in employment and in the provision of health care services. A disability is a condition that interferes with a major life function, such as blindness, paralysis, or coronary artery disease. More controversial are conditions that might make the person less likely to want to reproduce, bringing that person within the protection of the ADA since reproduction has been interpreted by courts to be a major life function. AIDS is such a condition, but so, too, might be some untreatable dominant single gene disorders, such as Huntington's disease. Courts' focus in interpreting the ADA has been on the severity of the disorder. Thus the ADA has application for complex multifactorial diseases where the manifestation is severe, such as coronary artery disease.

At the federal level, the Equal Employment Opportunities Commission has interpreted the Americans with Disabilities Act to cover individuals with genetic predispositions to later develop particular diseases. As people begin undergoing genetic testing for complex, common disorders, however, questions may be raised as to just what constitutes a disability under the ADA. For example, whether an alleged genetic predisposition to develop carpal tunnel syndrome should be considered a disability under the ADA is an issue before a federal court.

In the context of negligence, when people seek genetic testing, genetic counseling or other genetic information, health care providers have an obligation to provide it in a high quality way. When patients might benefit from genetic services, physicians have a legal obligation to offer them. Medical malpractice cases have held health care providers liable for not informing patients they were in a high-risk group with respect to certain genetic risk and for not performing genetic tests accurately.

In the negligence context, some courts have only allowed recovery if the disorder at issue was severe, but they have set a different standard for severity than in the ADA context. For example, one court has suggested that blindness would not be a sufficiently severe disability for parents to recover damages if the obstetrician failed to advise them of a genetic test to predict blindness or a laboratory failed to undertake the test accurately.

The rationale for finding physicians liable for negligence is that such liability deters low quality genetic services. However, the vast majority of these cases deal with single gene disorders such as Tay-Sachs disease or chromosomal abnormalities such as Down syndrome. The courts in the cases involving malpractice liability in the genetic testing arena have assumed that the test not offered or undertaken incorrectly was highly predictive.

The harm in the case was in not providing the patient with highly predictive genetic information. In one case, for example, a court refused to hold a physician liable for failing to offer a genetic test when the test would only have predicted 20% of the instances of the disorder. The court held, "A mere 20 percent chance does not establish a 'reasonably probable causal connection' between defendants' negligent failure to provide the [genetic] test and plaintiffs' injuries. A less than 50-50 possibility that defendants' omission caused the harm does not meet the requisite reasonable medical probability test of proximate cause." Yet, when dealing with complex, common disorders, a particular genetic test may not predict more than 20% of the cases. How then will low quality genetic services be deterred in the arena of complex, common disorders if such precedents are followed and courts refuse to find liability for failure to offer a test or failure to perform it correctly? Perhaps a new policy needs to be instituted in the negligence area that is closer to the ADA approach that focuses on the nature of the disorder rather than the nature of the genetic test for the disorder.

191. Genetic Materials: Resources, Rights, or Sacred Objects

Mervyn L. Tano
International Institute for Indigenous Resource Management
mervtano@iiirm.org

On June 4 and 5, 2001, the International Institute for Indigenous Resource Management convened a Roundtable on Genetic Materials: Resources, Rights, or Sacred Objects—Alternatives to the Intellectual Property Regime for Protecting Indigenous Genetic Materials in Denver, Colorado. The primary purpose of the Roundtable was to discuss alternative ways the genetic materials of indigenous peoples can be protected. The Roundtable participants were all well aware of the limitations of the current intellectual property regime so the major part of the two-day meeting was spent identifying alternative ways of characterizing human genetic materials and discussing alternative theories by which these genetic materials could be controlled by indigenous peoples, and the legal, anthropological and other research required to support these alternative theories. The poster will present the results of the roundtable.

192. The DNA Patent Database

Robert M. Cook-Deegan and LeRoy Walters
Kennedy Institute of Ethics, Georgetown University
bcd@nas.edu

Using an algorithm developed by James Martinell at the United States Patent and Trademark Office (USPTO), the coauthors, in collaboration with Stephen McCormack, have developed a method for tracking and reporting the number of human-DNA-based patents issued by USPTO. A database containing the numbers and texts of relevant patents is available at www.genomic.org. With the aid of this database it is possible to identify the organizations holding the largest numbers of DNA-based patents (the U.S. Government, the University of California, and Incyte are the top three organizations), as well as to follow trends in the numbers of patents issued (in 1996: 1,425; in 1997: 2,312; in 1998: 3,354; and in 1999: 4,250. Analyses based on this database are available at www.stanford.edu/class/siw198q/websites/genomics. The presentation will update the information on DNA-based patents through the end of the year 2001 and identify major trends and issues in this important field.

193. *Bioinformatics and the Human Genome Project*

Mark Bloom and Sherry Herron
Biological Sciences Curriculum Study
mbloom@bscs.org

Bioinformatics and the Human Genome Project is a curriculum module designed for high school biology classes. It specifically addresses how bioinformatics uses data generated by the Human Genome Project to help us understand how genes contribute to our health and well-being. Developed by BSCS, a nonprofit curriculum development group, the module uses an inquiry-based approach that also includes an analysis of related ELSI issues. The curriculum is the fifth genome module produced by BSCS and will be provided to teachers at no cost.

Students using the module should be familiar with Mendelian genetics, the chromosome theory of inheritance, the chemical nature of the gene (including the structure of DNA), and the central dogma, which states that genetic information resides in DNA, passes through an RNA intermediate, and is ultimately expressed as protein.

Designed for five periods of classroom instruction, the module includes extensive teacher background material, five student lessons, and an articulated Web site. The student lessons are organized into a conceptual whole that introduces basic techniques of bioinformatics and considers some of the ELSI issues related to informed consent and access to genetic data.

The first lesson creates an overall scenario for the module in which students work as employees of a bioinformatics company concerned with applying genomic data to the prevention, diagnosis, and treatment of cancer. Students first assemble, by hand, short DNA sequences. Small sequence differences illustrate genetic variation in the population and introduce the possibility of sequencing errors. Next, students look for open reading frames to see if their sequence may be part of a gene. Students then use the Web site to perform a BLAST search, which reveals the sequence to be part of the gene associated with ataxia telangiectasia (A-T). At this point, students change the course of their research from cancer to A-T. The final lesson deals with ethical issues related to informed consent and genetic privacy brought about by the new emphasis on A-T.

194. Damage Recognition, Protein Signaling, and Fidelity in Base Excision Repair

M. A. Kennedy[1], M. K. Bowman[1], G. W. Buchko[1], P. D. Ellis[1], J. H. Miller[1], D. F. Lowry[1], T. J. Straatsma[1], Susan S. Wallace[2], and David Wilson III[3]
[1]Battelle Pacific Northwest National Laboratory, Richland, WA 99352
[2]University of Vermont, Burlington, VT 05405
[3]Lawrence Livermore National Laboratory, Livermore, CA 95441
ma_kennedy@pnl.gov

Glycosylase studies: The base excision repair enzyme formamidopyrmidine-DNA glycosylas (FPG) from *E. coli* has been examined using solution-state NMR spectroscopy. Out of 252 possible amide proton-nitrogen correlations (269 residues - 16 prolines and the N-terminal amide), 209 amide cross peaks were observed (83%) and 180 of these were assigned (86%). Chemical shift perturbations have been observed for a subset of residues upon binding a 13 base pair DNA oligonucleotide containing a propane diol linker mimicking a nonhydrolyzable abasic site from which a chemical shift map of the DNA binding surface has been generated. Comparison of free-precession and CPMG HSQC data indicate the presence of slow millisecond to microsceond timescale motion for some residues in the absence of DNA that might be correlated with a hinge motion essential for DNA binding. Using similar analyses, attempts are being made to determine if DNA binding freezes the slow hinge motion observed in the absence of DNA.

APE1 studies: Histidine-aspartate (His-Asp) dyads are critical constituents in several key enzymatic reactions. The importance of these dyads is best exemplified in serine and trypsin-like proteases, where structural and biochemical NMR studies have revealed important pKa values and hydrogen-bonding patterns within the catalytic pocket. However, it has been debated whether the histidine in the His-Asp dyad can be partially charged, i.e. maintained at intermediate degrees of protonation. We used a new Oxford Instruments 21.1 Tesla superconducting magnet to measure the charge state of a critical histidine of apurinic endonuclease 1 (APE1), a human DNA repair enzyme that cleaves adjacent to abasic sites in DNA using an active site His-Asp pair. These observations imply that the role of the dyad in APE1 is not to directly bind a divalent cation as suggested. Rather, the dyad may activate a water molecule or bind DNA phosphate oxygen. EPR has also been used characterize the metal binding sites in APE1. Binding studies on APE-1 were investigated using oxovanadium, VO(II), as a surrogate probe for divalent metal binding sites. The CW-EPR spectra showed two different bound forms of VO(II) that could be explained either by coordination of three water and one hydroxyl as equatorial ligands to the oxovanadium or by three water and a histidine as ligands. Pulsed EPR spectra and 2D HYSCORE spectra showed water and histidine as direct ligands confirming the CW EPR analysis.

Pol b studies: Solid-state NMR is being used to characterize the metal environment in Pol b in relation to its catalytic function. We have refined our utilization of paramagnetic dopants to speed up the low temperature solid-state NMR experiment used to obtain magnesium spectra. We estimate that this has increased our data acquisition rates by a factor of 100. In addition, we have been examining the theory behind our experimental approach to in order to examine the potential of recovering Mg-Mg distance information (on the order of 3 to 5 Å). Such information will be critical to our understanding of the "relative rigidity" of the active site of Pol b during catalysis, i.e., does the active site remodel in the presence of either the correct or incorrect nucleotide triphosphate? Our initial results suggest we can obtain this information. Molecular dynamics simulations have been carried out for the human DNA Polymerase b with Gapped DNA, starting from the crystal structure reported by Sawaya et al. The reduced substrate ddCTP in the crystal structure has been replaced by dCTP, which is the correct nucleotide that is complementary to the template residue. This system was equilibrated for 400 ps, and a molecular dynamics simulation extending 1.5 ns has been completed. The molecular dynamics

simulation was carried out using the massively parallel NWChem computational chemistry software, with the AMBER force field and with particle-mesh Ewald (PME) long range electrostatic contributions.

195. Low Dose Ionizing Radiation-Induced Effects in Irradiated and Unirradiated Cells: Pathways Analysis in Support of Risk Assessment

Bruce E. Lehnert, Robert Cary, Donna Gadbois, and Goutam Gupta
Bioscience Division, Los Alamos National Laboratory, Los Alamos, New Mexico USA
lehnert@telomere.lanl.gov

We are investigating how human cells respond to low dose ionizing radiation (LDIR), i.e., <1-10 cGy, with emphasis on comparing and contrasting effects in directly irradiated cells and "bystander" cells. Since advanced risk assessment modeling and prediction ultimately require an in-depth understanding of the complexity of cellular responses to LDIR as they pertain to untoward or benign effects, we are undertaking a "systems biology" approach to obtain such information. The specific aims of the project are: 1) to assess the temporal changes in gene expression in directly irradiated and bystander cells as functions of LDIR dose, cell types, cell culture conditions, and subsequent LDIR exposure conditions, 2) relate gene expression profiles in directly irradiated and bystander cells with the corresponding temporal and spatial changes in expressed proteins, and 3) determine how LDIR-induced extracellular and intracellular oxidative/reductive landscapes, signal transduction pathways, and gene and protein expression profiles regulate cellular responses. Our central hypothesis is that differences in the gene expression profiles and temporal and spatial patterns of key proteins expressed in directly irradiated and bystander cells critically determine how the cells ultimately respond to LDIR. Consistent with DOE's new Genomes to Life initiative, knowledge gained from the project will contribute to the genesis of future hypothesis-driven investigations of candidate genes and proteins as they pertain to individual human resistance and susceptibility to the effects of LDIR, while additionally leading to the development of mechanistic models for predicting cancer risk.

196. The Application of Genome Data to the Important Problem of Risk from Low Dose Radiation

Antone L. Brooks
Washington State University, 2710 University Drive, Richland, WA 99352
tbrooks@tricity.wsu.edu

Efforts to sequence the human genome and to make this information available to the scientific community are already paying great dividends. New genomics data and technology make it possible to address important societal issues in biology, medicine, and even in health risk. One application has been to apply these techniques to determine the cellular and molecular responses induced by low doses of ionizing radiation. Before the genome project, it was not possible to determine biological responses to very low levels of ionizing radiation (below about 0.10 Gy). The Low Dose Radiation Research Program funded by the DOE Office of Biological and Environmental Research was made possible by the merging of new technological developments with the genome research. The overall goal of this program is to provide a sound scientific basis for radiation protection standards. The program has been in place for just over three years and is currently funding 54 projects. There have already been several major breakthroughs resulting in a re-evaluation of basic radiation paradigms on which current radiation risk standards were set. These breakthroughs are a direct result of the gene chip and sequencing technology generated by the genome program. There is now evidence that cells do not require a direct "hit" to exhibit changes in gene expression, gene mutation and chromosome damage, but may also respond if a neighbor cell is irradiated, a phenomenon called the "bystander" effects. Such observations make it necessary for us to re-evaluate the effective biological target size for radiation and the significance of the long held "hit theory" of radiation biology. It has also been demonstrated that exposure of the matrix on which cells grow can change both the pattern of gene expression and the cells phenotype to result in cell transformation without direct induction of mutations. Therefore, the relative role of mutations and gene expression in

cancer induction must be redefined. This may result in potential impacts on the basic linear-no-threshold hypothesis that is used in standard setting. Finally, low dose studies have demonstrated that the pattern and type of genes expressed after low doses of radiation are different from those observed after higher doses. Research has also shown that these patterns of gene expression influence many important genes involved in repair of DNA damage, as well as in programmed cell death (apoptosis). Results of recent studies suggest that low doses of radiation may decrease the level of spontaneous cell transformation resulting in another expression of the "adaptive response". Without the advances in genomics most of these observations would not have been possible. Their impact on radiation risk and standards remains to be determined. However, the research from the Low Dose Program will provide a sound scientific basis for radiation risks. Continued application of new equipment, methods and techniques will be important in addressing many important scientific and societal needs.

Research funded by US DOE Grant DEFG0399ER62787 to Washington State University

197. Genome-Scale Modeling of Low-Dose Irradiation Responses Using Microarray Based Gene Networks

Matthew Coleman[1], Terence Critchlow[1], Mike Colvin[1], Tom Slezak[1], David Nelson[1], and Leif Peterson[2]
[1]Molecular and Structural Biology Division, L-448, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California
[2]Departments of Medicine and Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, SCUR-924, Houston, Texas 77030
coleman16@llnl.gov

Cells and tissues with similar radiation response phenotypes are predicted to have common ionizing radiation (IR)-induced gene expression profiles that are controlled by shared groups of regulatory elements. Our overall objective is to utilize genome-scale expression microarray data in conjunction with DNA sequence/pattern databases to build a computer-based gene-network model for identifying, grouping and predicting regulatory elements that control differential aspects of the early cellular responses to IR. This research project will develop a prototype model by: (a) Grouping genes identified by microarray experiments into IR-responsive clusters based on their relative-transcript abundance and their differential IR radiation responses at low (10cGy) and high doses and (b) Identifying regulatory elements (and their locations relative to the open reading frame) that distinguish among separate IR responsive gene clusters. To provide initial evaluation of our model, we will determine whether the identified regulatory elements are conserved across species and valid for microarray IR-responsive data sets from other laboratories. This model will be expanded and refined by including additional microarray expression data (low versus high dose responses, early versus late responses, adaptive response, and tissue/cell differences) and additional sequence data as they become available. A validated model will allow us to identify new human genes likely to be IR modulated and identify genes/pathways that are associated with different radiation response phenotypes (e.g., low dose sensitivity, adaptive response, sensitivity to chromosome damage, etc.) The identification and characterization of regulatory element profiles of IR-responsive genes will provide valuable understanding of the genetic mechanisms of IR-response and should provide powerful biological indicators of genetic susceptibilities for tissue and genetic damage.

198. Molecular Mechanisms and Cellular Consequences of Low-Dose Exposure to DNA-Damaging Agents

Andrew J. Wyrobek[1], Matthew Coleman[1], Eric Yin[1], Francesco Marchetti[1], Sandra McCutchen-Maloney[1], Allen Christian[1], David Nelson[1], Irene Jones[1], Larry Thompson[1], Leif Peterson[2], and Jian-Jian Li[3]
[1]Molecular and Structural Biology Division, L-448, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550

[2]Departments of Medicine and Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, SCUR-924, Houston, Texas 77030
[3]Department of Radiation Research, Beckman Research Institute, City of Hope National Medical Center, Duarte, CA 91010-3000
wyrobek1@llnl.gov

It is well established that high-dose exposures to ionizing radiation (IR) leads to diverse pathologies of skin and other tissues and to serious late-onset diseases including cancers. The long-term objectives of this research are to investigate the early cellular effects of low-dose exposures (1-10cGy) in human and mouse cells and to determine whether the early changes in gene expression (mRNA or protein) of specific gene/pathways are associated with subsequent risk for cytogenetic damage. In this ongoing project, we utilized both LLNL-manufactured cDNA arrays and commercial oligonucleotide arrays to identify hundreds of low-dose (10cGy) responsive genes in irradiated human lymphoblastoid (HLB) cell lines and mouse brain. We also identified a group of human genes that show transcriptional changes associated with adaptive responses (AR) and identified several human protein peaks by SELDI mass-spectrometry. In parallel experiments in the mouse, we established baselime tissue variations in the expressions of DNA repair and stress response genes, identified hundreds of genes whose expression was modulated after 10cGy brain irradiation, and studied Rad54 KO mice to suggest that double strand break (DSB) repair is associated with altered expression response. The renewal project will: (a) use genome-scale microarrays to survey the human and mouse genome for low dose (1-10cGy) radiation-responsive genes (b) investigate the nature of dose- and time-response profiles across mouse tissues, human lymphoblastoid (HLB) cells, and fresh human lymphocytes; (c) identify and characterize genes whose changes in early expression are associated with adaptive response, beginning with candidate genes already identified by their mRNA and the protein levels; and (d) continue our studies in mice of tissue variations and the role of DSB repair in radiation response. This project will provide new knowledge of the early cellular responses to low-dose IR to reduce the uncertainty of assessing risk at low-dose levels. It is also expected to identify genes whose expression is associated with low dose IR exposure and susceptibility for adaptive response.

199. Molecular Mechanism of the 9-1-1 Checkpoint Response to DNA Damage Based on Protein Structure Prediction

Ceslovas Venclovas, Michael Colvin, and
Michael P. Thelen
Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory
mthelen@llnl.gov

Deciphering the molecular mechanisms of DNA replication and repair is hindered by the lack of detailed structural information for the protein machinery involved. For structurally uncharacterized proteins, modeling by comparison can generate hypotheses for fundamental DNA repair mechanisms, serving as a powerful guide for experimental design. Here we describe such a three-dimensional structure prediction for two interacting complexes of nine proteins that are central to the DNA damage checkpoint in eukaryotic cells.

The checkpoint response to DNA damage requires Replication Factor C (RFC), a hetero-pentameric protein complex that is essential to eukaryotic replication. Current experimental evidence indicates that during normal DNA replication, RFC binds to primed DNA and uses ATP to drive the loading of PCNA, the sliding clamp that tethers polymerase to DNA for processive DNA synthesis. Upon DNA damage, Rad17 complexes with RFC and causes it to load a different sliding clamp protein, the Rad9-Rad1-Hus1 (9-1-1) heterotrimer. Based on our comparative modeling studies and more recent experimental observations, the 9-1-1 complex behaves similarly to PCNA, as a polymerase processivity/fidelity factor, but is specifically needed when the DNA template contains breaks or adducts that would halt normal replication.

Five distinct RFC subunits and Rad17 have detectable sequence similarity to each other and to functionally analogous proteins from bacteria and archaea. We used recently determined structures, including the RFC small subunit from *Pyrococcus furiosus*, to produce high confidence, all-atom models for each RFC subunit. The quaternary structure of the RFC complex was then assembled by analogy to the *E. coli* clamp loader, or gamma complex. Constraints derived from available

structural, biochemical and genetic data were used to predict relative positions for the individual Rad17 and RFC subunits within the complex. The resulting architectural model of Rad17/RFC includes interactions with the 9-1-1 ring, and with DNA and ATP, leading to a first approximation of the eukaryotic mechanism for dealing with DNA damage during the cell division cycle. The model indicates specific protein-protein contact regions that could be tested by site-directed mutation.

Work performed under auspices of U. S. Department of Energy by the University of California, LLNL under Contract No. W-7405-Eng-48

200. Phylogenetic Analysis of Two Human Proteins that are Homologues of Proteins Involved in Base Excision Repair, Formamidopyrimidine DNA Glycosylase and Endonuclease VIII

Sirisha Sunkara, Susan S. Wallace, and **Jeffrey P. Bond**
Department of Microbiology and Molecular Genetics, Markey Center for Molecular Genetics, University of Vermont, Stafford Hall, Burlington, VT 05405-0068, USA
Jeffrey.Bond@uvm.edu

It is important to identify human proteins involved in DNA repair because understanding DNA repair is essential for understanding carcinogenesis. We report identification of two human proteins that are members of a family of proteins involved in base-excision repair, the formamidopyrimidine DNA glycosylase (Fpg)/endonuclease VIII (Nei) family. Phylogenetic analysis suggests that an ancestor of plants, fungi, and metazoa possessed Fpg and that a monophyletic group of Nei sequences exists. On the basis of phylogenetic analysis we classify one of the sequences as Fpg and the other as Nei. Analysis of alignments of the sequences of the human proteins with previously identified Fpg and Nei sequences in the context of the structure of a bacterial Fpg suggests that the human proteins have structures similar to that of the bacterial Fpg. In particular, sequence similarities between the human sequences and previously identified Fpg/Nei sequences include highly conserved regions associated with each of two Fpg structural domains and amino acids involved in catalysis. Eukaryotic Fpg sequences are missing the zinc finger motif seen in bacterial Fpg. Human Nei has a C-terminal extension that includes a second zinc finger motif and a region similar to portions of certain AP endonuclease and topoisomerase sequences.

Infrastructure

201. HGMIS: Making Genome Science and Implications Accessible

Anne E. Adamson, Jennifer L. Bownas, **Denise K. Casey**, Sherry A. Estes, Sheryl A. Martin, **Marissa D. Mills**, Judy M. Wyrick, Laura N. Yust, and **Betty K. Mansfield**
Life Sciences Division; Oak Ridge National Laboratory; 1060 Commerce Park, MS 6480; Oak Ridge, TN 37830
mansfieldbk@ornl.gov
www.ornl.gov/hgmis/
DOEGenomesToLife.org

Working with scientific teams to communicate BER's programs to the scientific community and the public to help the Department of Energy fulfill its broad missions in energy, environmental remediation, and the protection of human health.

The multidisciplinary Human Genome Project (HGP) has revolutionized life sciences research and applications so profoundly that this century has been dubbed the "biology century." Applications of knowledge and technologies derived from the genomics era that began in the late 20th century will affect almost everyone. Entirely new approaches are being implemented to the practice of medicine and agriculture and to biological research, including the new DOE program Genomes to Life. For an unprecedented understanding of the inner workings of whole biological systems, genetic data will provide the foundation upon which research from many biological subdisciplines will be layered.

Since 1989 the Human Genome Management Information System (HGMIS) has been producing and distributing information about the HGP and related science to (1) facilitate research progress and (2) increase public understanding and the accessibility of genome science and its societal implications so that more-informed personal and public policy decisions can be made. Democratizing access to genetic science should help maximize the benefits while protecting against misuse of the data.

HGMIS products include the technical newsletter *Human Genome News*, progress reports, fact sheets, invited articles in peer-reviewed publications, posters, and primers. Using knowledge-management experience gained in this work and in presentations, exhibitions, and judicial and minority meetings, HGMIS initiated and is continually developing and expanding a suite of Web sites for a variety of audiences. Because genomics and the life sciences are becoming so pervasive in all sectors of society, HGMIS seeks to make information accessible to nontechnical audiences as well as to scientists, social scientists, and medical and legal practitioners; these professionals, particularly, need an understanding of genetics to enhance their work and allow them to communicate across disciplines. Containing a unique compilation of resources not found in any discipline-specific publication, *Human Genome News*, for example, is distributed to some 20,000 print subscribers and many more via the Web. Each month, more than 400 information packages are mailed out and numerous e-mail and phone queries are answered directly.

Our suite of Web sites, comprising the Human Genome Project Information (HGPI), Microbial Genome Program, Genomes to Life, and CERN Virtual Library on Genetics, now supports some 275,000 unique user sessions and 900,000 text-file and 8.5 million total-file transfers each month. The sites contain more than 4200 files, some 3500 of which are text files. The average length of time per visit is 12.5 minutes. About 8000 other sites link to HGPI and its individual pages. Although these DOE-sponsored Web pages are presented from the genomics research perspective, they extend well beyond the project's primary goals into the technological and societal ramifications of genomics research. They are important resources for nearly all major news and science news outlets carrying stories on genomics, including CNN, MSNBC, ABC, CBS, Yahoo, *Wired*, *Nature*, and *Science*.

HGMIS Web pages are updated daily, entirely new pages are added several times each year, and all sites receive a major overhaul at least annually. HGMIS continually incorporates feedback from Web users into the strategy for the sites' ongoing development. Main priorities are to meet user needs for accurate, understandable, and easy-to-locate information

relevant to genomics and related science (including new DOE programs) and their societal implications.

HGMIS publications and Web pages have won numerous endorsements from such sites as Schoolsnet.com, BigChalk.com, KidsHealth.org, CyberU.com, sciLINKS.org, Geniusfind.com, Hardin MD Clean Bill of Health, ISI Current Web Contents, and Awesome Library. Awards include First Place in Online Communications in the 2001 regional and international competitions sponsored by the Society for Technical Communication. The *DOE Primer on Molecular Genetics* was voted one of *Scientific American*'s top 50 Web sites for 2001, and *Science* magazine's Web site named several HGMIS Web pages as "possibly the best single resource for those new to genetics and genomics."

HGMIS has assisted numerous organizations and is actively collaborating with others, including Qiagen, Inc.; the *Journal for Minority Medical Students*; EurekAlert!; the American Museum of Natural History; and the Zeta Phi Beta Sorority, Inc.

Constructive comments are appreciated.

This work is sponsored at Oak Ridge National Laboratory by the Office of Biological and Environmental Research, U.S. Department of Energy, under contract No. DE-AC05-00OR22725 with UT-Battelle LLC.

Appendices

Appendix A: Author Index

Contact authors are in **bold**.

M

Appendix B: National Laboratory Index

U.S. Department of Energy Laboratories

Genomic research at DOE laboratories is described on the following pages.