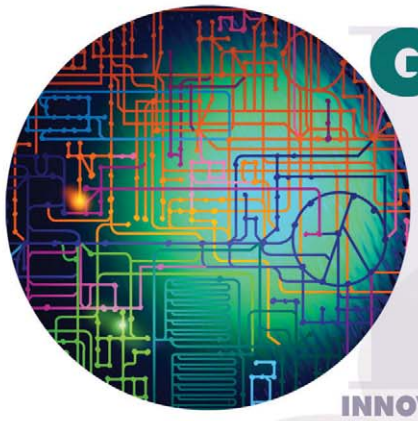
DOE/SC-0072



GENOMES to LIFE

BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES

U.S. DEPARTMENT OF ENERGY

INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS

Contractor-Grantee Workshop I

Arlington, Virginia

February 9–12, 2003

Office of Biological and Environmental Research
Office of Advanced Scientific Computing Research

Office of Science
U.S. DEPARTMENT OF ENERGY

Genomes to Life Program

Gary Johnson
U.S. Department of Energy (SC-30)
Office of Advanced Scientific Computing Research
301/903-5800, Fax: 301/903-7774
gary.johnson@science.doe.gov

Marvin Frazier
U.S. Department of Energy (SC-72)
Office of Biological and Environmental Research
301/903-5468, Fax: 301/903-8521
marvin.frazier@science.doe.gov

A limited number of print copies are available. Contact:

Sheryl Martin
Oak Ridge National Laboratory
1060 Commerce Park, MS 6480
Oak Ridge, TN 37830
865/576-6669, Fax: 865/574-9888, *martinsa@ornl.gov*

An electronic version of this document became available on February 4, 2003, at the Genomes to Life Web site:

- *http://doegenomestolife.org/pubs/2003abstracts/*

Abstracts for this publication were submitted via the Web.

DOE/SC-0072

GENOMES to LIFE

BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES

U.S. DEPARTMENT OF ENERGY

INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS

Contractor-Grantee Workshop I

Arlington, Virginia

February 9–12, 2003

Prepared for the
U.S. Department of Energy
Office of Science
Office of Biological and Environmental Research
Office of Advanced Scientific Computing Research
Germantown, MD 20874-1290

Prepared by
Human Genome Management Information System
Oak Ridge National Laboratory
Oak Ridge, TN 37830
Managed by UT-Battelle, LLC
For the U.S. Department of Energy
Under contract DE-AC05-00OR22725

Contents

Session and poster board numbers are indicated in the gray boxes.

Sandia National Laboratories

Session and poster board numbers are indicated in the gray boxes.

Session and poster board numbers are indicated in the gray boxes.

Session and poster board numbers are indicated in the gray boxes.

Session and poster board numbers are indicated in the gray boxes.

Session and poster board numbers are indicated in the gray boxes.

Session and poster board numbers are indicated in the gray boxes.

Total Number of Abstracts: 64

Session and poster board numbers are indicated in the gray boxes.

Welcome to Genomes to Life Contractor-Grantee Workshop I

Welcome to the first of what we hope will be many Genomes to Life (GTL) contractor-grantee workshops. Although only in its second official year of funding, GTL already is attracting broad and enthusiastic interest and support from scientists at universities, national laboratories, and industry; colleagues at other federal agencies; Department of Energy leadership; and Congress.

You are part of an exciting era in biology as we begin to systematically leverage the knowledge and capabilities brought to us by DNA sequencing projects into an understanding of the functioning and control of entire biological systems. GTL certainly is not the first, nor will it be the last, to conduct "systems biology" research, but we believe the program offers a roadmap for these new explorations. GTL research is, of necessity, at the interface of the physical, computational, and biological sciences.

GTL will require the development of technologies that will enable us to "see" biology happen at finer scales of resolution. It also will require a substantial integration of our broad capabilities in mathematics and computation with our new knowledge of biology. Only with this integration can we achieve GTL's fundamental goal: to understand biological systems so well that we can accurately predict their behavior with sophisticated computational models.

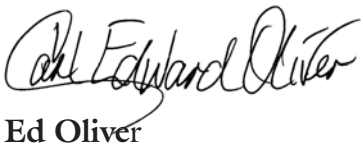
To enable this goal, GTL aims to develop these new technological, analytical, biological, and computational capabilities into cost-effective, widely accessible, high-throughput capabilities analogous to today's DNA sequencing factories.

Microbes are GTL's principal biological focus. In the complex "simplicity" of microbes, we find capabilities needed by DOE—indeed by our entire nation—for clean energy, cleanup of environmental contamination, and sequestration of atmospheric carbon dioxide that contributes to global warming. In addition, the fundamental knowledge and technologies developed in GTL will be broadly usable in all areas of biological research.

This first GTL program workshop is an opportunity for all of us to discuss, listen, and learn about the exciting science, identify research needs and opportunities, form research partnerships, and share the excitement of this program with the broader scientific community.

We look forward to a stimulating and productive meeting and offer our sincere thanks to all the organizers and to you, the scientists, whose vision and efforts will help us all to realize the promise of this exciting research program.

Ari Patrinos
Associate Director for
Biological and Environmental Research
Office of Science
U.S. Department of Energy



Ed Oliver
Associate Director for
Advanced Scientific Computing Research
Office of Science
U.S. Department of Energy

DOE/SC-0069

Genomes to Life: Realizing the Potential of the Genome Revolution

Office of Science
January 2003
DOEGenomesToLife.org

The remarkable successes of the Human Genome Project and spin-offs revealing the details of numerous genomes—from microbes to plants to mice—provide the richest resource in the history of biology. These achievements now empower scientists to address the ultimate goal of modern biology: to obtain a fundamental, comprehensive, and systematic understanding of life. This goal is founded, as is life itself, on the genome, whose genes encode the proteins that carry out most cellular activities via a labyrinth of pathways and networks that make the cell "come alive" (see figure below).

Catalyzing Systems Biology

The Department of Energy's (DOE) Genomes to Life (GTL) program is combining high-throughput advanced technologies and computation with the information found in microbial genomes to establish a foundation for achieving an understanding of living systems (see "Microbes for DOE Missions," p. 2). GTL is designed to help launch biology onto a new trajectory to comprehensively understand cellular processes in a realistic context. This new level of exploration, known as systems biology, will empower scientists to pursue completely new approaches to discovery, transforming biology to a more quantitative and predictive science. GTL scientific goals target the fundamental processes of living systems by studying them on three levels:

1. Proteins and multicomponent molecular machines that form all of the cell's structures and perform most of the cell's work.
2. Gene regulatory networks and pathways that control cellular processes.
3. Microbial communities in which groups of cells carry out complex processes in nature.

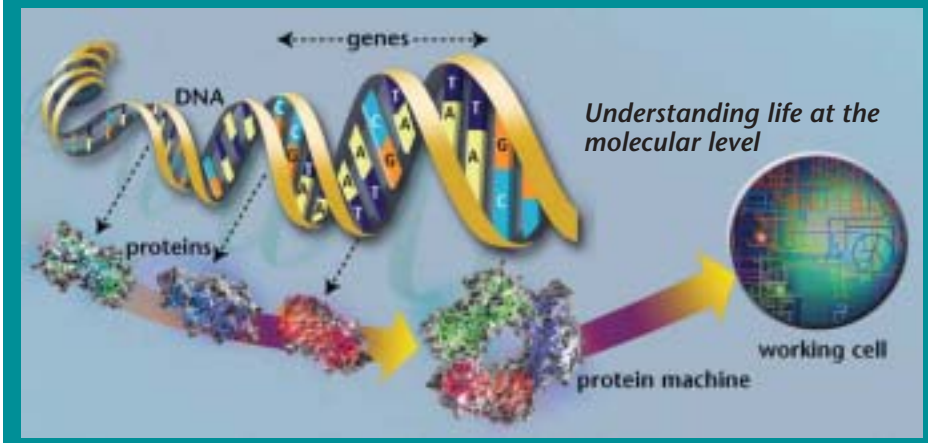
Molecular machines carry out chemical reactions, generate mechanical forces, transport metabolites and ions, and make possible every action of a biological system. A cell does not generate its entire repertoire of molecular machines at once. Genomic regulatory elements dictate the particular set produced according to the organism's life strategy and in response to environmental cues, including other microbial populations in the larger ecological community.

A comprehensive approach to understanding biological systems thus extends from individual cells to many cells functioning in communities. Such studies must encompass proteins, molecular machines, pathways, networks, cells, and, ultimately, their regulatory elements, cellular systems, and environments. This next generation of biology is viable only with vastly increased computational and informational capabilities to master the full complexities of biological systems.

Understanding life processes at the molecular level is a "national science priority."

—OSTP-OMB FY 2004 budget guidance memo; see p. 6.

Genomes to Life: From DNA Sequence to Living Systems



Genes are made up of DNA and contain the information used by other cellular components (e.g., RNA and ribosomes, not shown here) to create proteins. A working cell is tightly packed with tens of thousands of proteins and other molecules, often working together as multimolecular "machines" to perform essential cellular activities (see also cell figure, p. 5).

Just as DNA sequencing capability was completely inadequate at the beginning of the Human Genome Project (HGP), the quantity and complexity of data that must be collected and analyzed for systems biology research far exceed current capabilities and capacities. The HGP taught that aspects of biological research can be made high-throughput and cost-effective (see graph, p. 5). Collecting and using such data and reagents will require a new organizational model that coordinates and integrates dozens of high-throughput technologies and approaches, some not yet refined or even developed. This is the central principle of GTL and indeed of all systems biology research.

Analysis of living systems will require a new generation of experimentation and the computational methods and capabilities to assimilate, understand, and model the data on the scale and complexity of real living systems. Computing must guide the research questions and interpretation at every step.

The knowledge base resulting from the GTL program will provide the entire research community with data, models, and simulations of gene expression, pathways, and network systems; molecular machines; and cell and community processes. These new capabilities and resources will inspire revolutionary solutions to DOE mission needs and transform the entire life sciences landscape, from agriculture to human health.

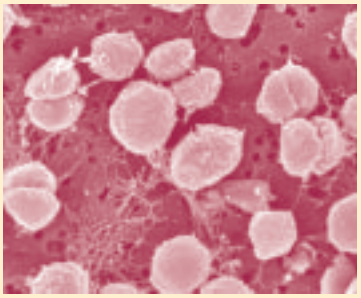
Microbes for DOE Missions: Energy Security, Cleanup, Climate Change

Why Study Microbes?

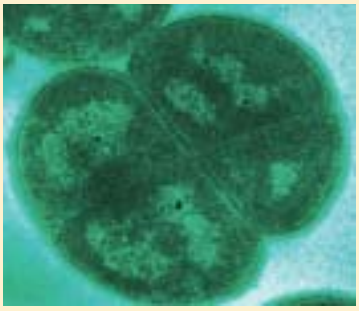
The ability of this planet to sustain life is largely dependent on microbes. They are the foundation of the biosphere, controlling earth's biogeochemical cycles and affecting the productivity of the soil, quality of water, and global climate. As one of the most exciting frontiers in biology today, microbial research is revealing the hidden architecture of life and the dynamic, life-sustaining processes on earth. The diversity and range of their environmental adaptations mean that microbes long ago "solved" many problems for which scientists are seeking solutions today (see examples at right). The incomprehensible number of microbes is an untapped but valuable resource that ultimately may be used to generate new energy sources (e.g., hydrogen for a new energy economy), new cleanup and industrial processes, and new ways of using biology to address DOE missions.

The Challenge

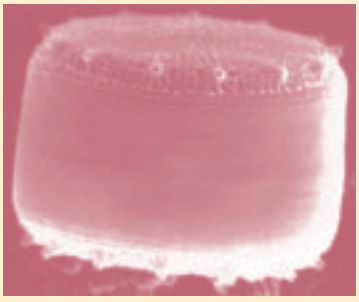
Microbes have become masters at living in almost every environment, harvesting energy in almost any form. Their sophisticated biochemical capabilities can be utilized for transforming wastes and organic matter, cycling nutrients, and, as part of the photosynthetic process, converting sunlight into energy and "fixing" (storing) CO_2 from the atmosphere. The analytical complexity involved in understanding these processes is enormous. Thousands of microbes have capabilities of interest. Moreover, each microbial genome contains thousands of genes capable of producing an even-greater number of proteins. These proteins combine to form innumerable molecular machines in myriad pathways and networks, many of which carry out biological processes useful for DOE missions (see "Potential Applications of GTL Science," p. 3).



Methanococcus jannaschii: Produces methane, an important energy source; contains enzymes that withstand high temperatures and pressures; possibly useful for industrial processes.



Deinococcus radiodurans: Survives extremely high levels of radiation and has high potential for radioactive waste cleanup.



Thalassiosira pseudonana: Ocean diatom that is major participant in biological pumping of carbon to ocean depths and has potential for mitigating global climate change.

Emerging Technologies and Computing for Systems Biology: Establishing a Firm Foundation for Genomes to Life

Numerous projects funded by the Office of Biological and Environmental Research (BER) over the past 5 years have established a strong foundation for the GTL program. These projects underscore the need for high-throughput biological research and novel computational approaches. They are also demonstrating the power of mass spectrometric analyses of whole microbial proteomes, developing new imaging methods, advancing the use of microarrays for expression analyses, exploring scalable ways to generate microbial proteins, and developing computational tools for second-generation genome analysis and annotation.

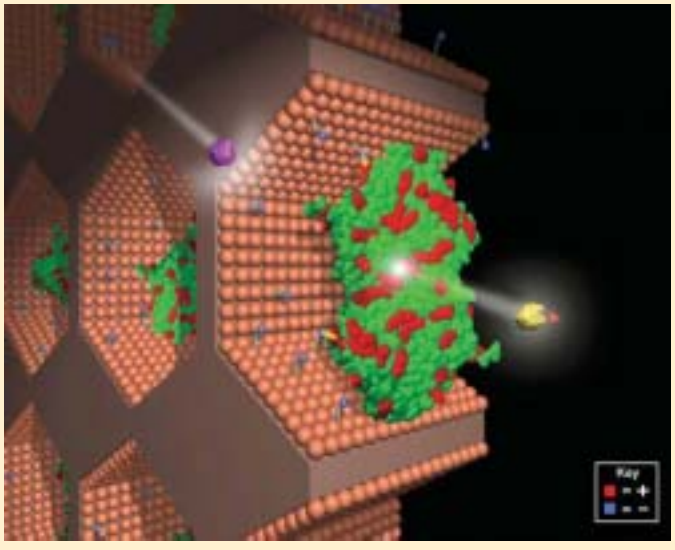
Several collaborative groups are integrating technologies and computational modeling to gain a systems understanding of specific microbes in their natural environments. For example, the *Shewanella* Federation, consisting of teams from academia, national laboratories, and other organizations is making progress in preliminary proteome and expression analyses of this remarkably versatile organism that can immobilize toxic uranium from ground water. By focusing multiple technologies on a single organism, the federation is integrating diverse experimental results into a multidimensional perspective of the biology of this key microbe. Thus far, the group has identified >77% of the predicted proteome of *Shewanella* (3782 of 4855 predicted genes) using ultrahigh-resolution mass spectrometry techniques. This and other groundbreaking BER projects (e.g., on *Deinococcus radiodurans*) have elucidated the highest percentages of the proteomes of organisms studied to date. These projects have also set the stage and identified the need for developing high-throughput user facilities accessible to the biological research community (see p. 5).

Office of Science—At the Forefront of the Biological Revolution

In 1986 the DOE Office of Science launched the Human Genome Project to understand, at the DNA level, the effects of energy production on human health. The HGP's innovative operational model proved highly successful, and benefits far exceeded the original goal. Today, DOE is poised to take the next vital steps—translating the genetic code in DNA into a new understanding of how life works and applying those biological processes to serve its challenging missions. Effective use of microbial and other biological systems and components will generate new biotechnological industries involving fuels, biochemical processing, nanomaterials, and broader environmental and biomedical applications.

The Office of Science has the capabilities and institutional traditions to bring the biological, physical, and computing sciences together at the scale and complexity required for success. Its academic affiliations, national laboratories, and other resources include major facilities for DNA sequencing and molecular-structure characterization, the high-performance computing resources of the Office of Advanced Scientific Computing Research (OASCR), the expertise and infrastructure for technology development, and a tradition of productive multidisciplinary research essential for such an ambitious research program. In the effort to understand biological systems, these strong assets and the GTL program will complement and extend the capabilities and research efforts supported by the National Institutes of Health, National Science Foundation, other agencies and institutions, and industry.

Potential Applications of GTL Science

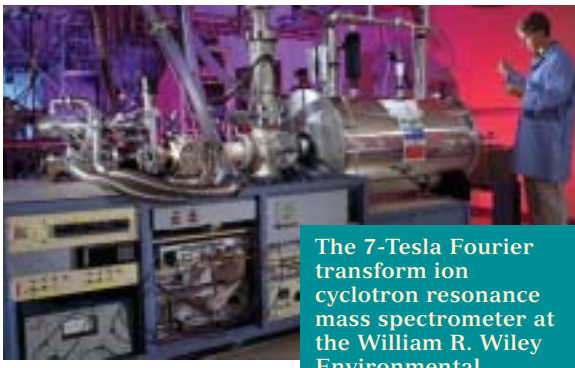


Learning about the inner workings of microbes and their diverse inventory of molecular machines can lead to discovery of ways to isolate and use these components to develop new, synthetic nanostructures that carry out some of the functions of living cells. In this figure, the enzyme organophosphorus hydrolase (OPH) has been embedded in a synthetic nanomembrane (mesoporous silica) that enhances its activity and stability [*J. Am. Chem. Soc.* 124, 11242–43 (2002)]. The OPH transforms toxic substances (purple molecule at left of OPH) to harmless byproducts (yellow and red molecules at right). Applications such as this could enable development of efficient enzyme-based ways to produce energy, remove or inactivate contaminants, and sequester carbon to mitigate global climate change. The knowledge gained from GTL research also could be highly useful in food processing, pharmaceuticals, separations, and the production of industrial chemicals.

Emerging Technologies and Computing for Systems Biology: Genomes to Life FY 2002 Awards

Genomes to Life continues to build its R&D portfolio, having made awards in July 2002 that totaled $103 million for FY 2002–FY 2007. These projects are focusing on isolating and characterizing protein machines, understanding complex biological communities, modeling cellular metabolic and regulatory processes, and modeling carbon sequestration processes in marine microbes.

Projects were chosen to test the concept of systems biology applications and to demonstrate advanced technologies (see picture at right), computation, and potential high-throughput methods in areas having possible impact on DOE missions. These awards represent the culmination of nearly 3 years of planning by the DOE Office of Science and hundreds of scientists at universities, national laboratories, and industry. The microbes studied in the pilot projects, as well as the 2002 awards, have potential for bioremediating metals and radionuclides, degrading organic pollutants, producing energy feedstocks including biomass and hydrogen, sequestering carbon, and demonstrating importance in ocean carbon cycling.



The 7-Tesla Fourier transform ion cyclotron resonance mass spectrometer at the William R. Wiley Environmental Molecular Sciences Laboratory. Mass spectrometry is the most sensitive method for identifying proteins.

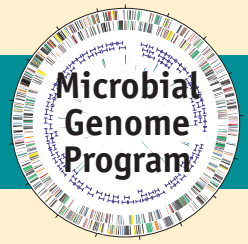
Institutions and Projects Awarded in 2002

- *Oak Ridge National Laboratory:* Developing technologies needed to identify and characterize the complete set of multiprotein complexes within a microbe involved in the carbon cycle (important for carbon sequestration) and another microbe that has the ability to clean up metals in contaminated soil (www.ornl.gov/GenomestoLife/).
- *Lawrence Berkeley National Laboratory:* Developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to environmental conditions found at DOE waste sites (vimss.lbl.gov/).
- *Sandia National Laboratories:* Developing experimental and computational methods to understand the proteins, protein-protein interactions, and gene regulatory networks in a marine microbe that plays a significant role in earth's carbon cycle; important for carbon sequestration (www.genomes-to-life.org/).
- *University of Massachusetts, Amherst:* Developing computational models to predict the activity of natural communities of microbes having potential for uranium bioremediation and production of electricity through their ability to transfer electrons to electrodes (DOEGenomesToLife.org/research/umass.html).
- *Harvard Medical School:* Studying the proteins, protein-protein interactions, gene regulatory networks, and community behavior of microbes active in the carbon cycle (with capabilities relevant to carbon sequestration); important for bioremediation strategies. Developing computational methods to understand the complex biology of these microbes at a systems level (arep.med.harvard.edu/DOEGTL/).

Other Participating Institutions

Argonne National Laboratory
Brigham and Women's Hospital
Diversa Corporation
Los Alamos National Laboratory
Massachusetts General Hospital
Massachusetts Institute of Technology
National Center for Genome Resources
Pacific Northwest National Laboratory
The Institute for Genomic Research
The Molecular Science Institute
University of California (Berkeley, San Diego, Santa Barbara)
University of Illinois
University of Michigan
University of Missouri
University of North Carolina
University of Tennessee (Knoxville, Memphis)
University of Utah
University of Washington

Microbes studied in GTL have had their genetic sequences determined under DOE's Microbial Genome Program.



Direct Web Access

- doegenomestolife.org/pubs.html
- doegenomestolife.org/gallery/images.html
- doegenomestolife.org/research/index.html
- www.ornl.gov/microbialgenomes
- www.ornl.gov/hgmis/education/education.html

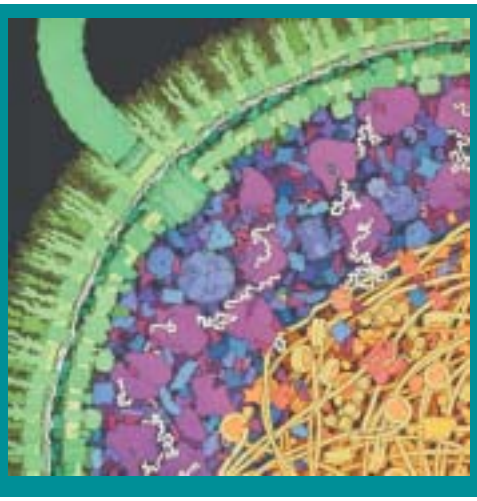
FY 2003 Call for Proposals: www.er.doe.gov/production/grants/Fr03-05.html

Analyzing whole microbial systems requires economies of scale. Traditionally, scientists have tried to understand the functions of individual proteins or small groups of proteins. In the new era of systems biology, researchers will study the behavior of the cell's entire working complements of proteins (proteomes), their regulatory pathways, and their interactions as they perform functions. These activities must be carried out on a scale that far exceeds today's capacities.

To meet this challenge, BER and OASCR have planned a set of four core research facilities. Building on each other, these facilities are intricately linked in their long-term goals, targets, technologies, capabilities, and capacities. They will provide scientists with an enduring comprehensive ability to understand and, ultimately, reap enormous benefit from the biochemical functionality of microbial systems. Making the most advanced technologies and computing resources available to scientists in large or small laboratories will democratize access to the tools needed for systems biology. They thus open new avenues of inquiry, fundamentally changing the course of biological research and greatly accelerating the pace of discovery. Hallmarks of these facilities include high-throughput advanced technologies, automation, and tools for data management and analysis, simulation, and an integrated knowledge base.

Large-Scale Facilities Spur Cost, Productivity Improvements

The dramatically increased productivity and reduced costs achieved in the HGP via high-throughput production environments (e.g., the DOE Joint Genome Institute) provide the paradigm for the dedicated industrial-scale facilities envisioned for Genomes to Life.



Crowded cell by D. Goodsell (©1999)

Large-scale facilities are required for identifying, characterizing, and modeling the activity of the tens of thousands of interacting components present at any given moment in a cell.

A Plan for GTL User Facilities

Facility I for Production and Characterization of Proteins would use highly automated processes to mass-produce and characterize proteins directly from microbial genome data and create affinity reagents ("tags") to identify, capture, and monitor the proteins from living systems.

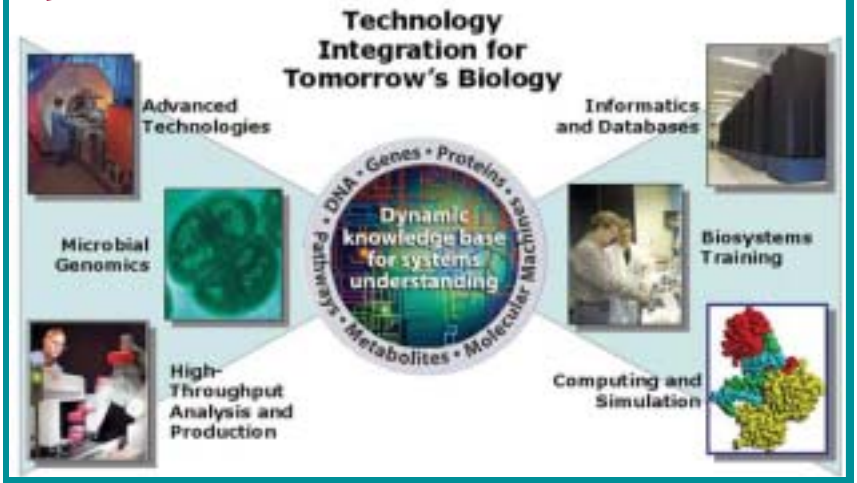
Facility II for Whole Proteome Analysis would characterize the expressed proteomes of diverse microbes under different environmental conditions as an essential step toward determining the functions and interactions of individual proteins and sets of proteins.

Facility III for Characterization and Imaging of Molecular Machines would isolate, identify, and characterize thousands of molecular machines from microbes and develop the ability to image component proteins within complexes and to validate the presence of the complexes within cells.

Facility IV for Analysis and Modeling of Cellular Systems would combine advanced computational, analytical, and experimental capabilities for the integrated observation, measurement, and analysis of spatial and temporal variations in the structures and functions of cellular systems—from individual microbial cells to complex communities and multicellular organisms.

GTL User Facility Hallmarks

Open Access to Data and Facilities



These facilities would serve as focal points for the life sciences research community, providing a national venue to pursue multidisciplinary systems biology and promote cross-disciplinary education.

OSTP, OMB: "National Science Priority"

Achieving a molecular-level understanding of life processes is a national science priority, according to the Office of Science and Technology Policy (OSTP) and Office of Management and Budget (OMB) *FY 2004 Interagency Research and Development Priorities*. The view in this guidance reflects that found throughout much of the biological research community.

AAM: "Develop New Technologies"

Specific recommendations made by the American Academy of Microbiology (AAM) in its 2001 colloquium report, *Microbial Ecology and Genomics: A Crossroads of Opportunity* include the following:

- "Develop new technologies including methods for measuring the activity of microorganisms (at the level of populations and single cells), approaches to cultivating currently uncultivable species, and methods for rapid determination of key physiological traits and activities.
- "Establish mechanisms to encourage the necessary instrument development.
- "Encourage instrumentation development through collaboration with device engineers, chemists, physicists, and computational scientists, since uncovering the diversity and activities of the microbial world is dependent on such advances.
- "Develop technology and analysis capability to study microbial communities and symbioses holistically, measuring system-wide expression patterns (mRNA and protein) and activity measurements at the level of populations and single cells."

BERAC Subcommittee: "Create Unique, High-Throughput Research Facilities"

The BER program of DOE, having played a critical, catalytic role in bringing about the genomic revolution, is now poised to make equally seminal contributions to the next, transforming phase of biology. A subcommittee report approved by the BER Advisory Committee (BERAC) in April 2002 stated: "DOE should now create unique, high-throughput research facilities and resources to translate the new biology, embodied in the Genomes to Life (GTL) program, into a reality for the nation. . . . [GTL] is designed to build on the major accomplishments of the past decade and move from this vision to reality—to a new and comprehensive systems approach from which we will understand the functioning of cells and organisms and their interactions with their environments. Since the science has changed so profoundly, to accomplish these challenging goals in a timely and cost-effective fashion, new facilities and new scientific resources are needed."

GTL Program Development

Genomes to Life is a joint program of the Office of Biological and Environmental Research and the Office of Advanced Scientific Computing Research in the Office of Science of the U.S. Department of Energy.

To solicit guidance from the scientific community in the development of the GTL program, in the past 2 years DOE has sponsored 15 workshops, which were attended by scientists from industry, national laboratories, and academia. A strategic plan for developing advanced and high-throughput facilities to serve GTL and the entire community was approved by the BER Advisory Committee (BERAC) in April 2002, and BERAC voiced approval of the subsequent facilities plan in December 2002.

GTL was developed in response to a 1999 charge by the DOE Office of Science to BERAC to define DOE's potential roles in post-HGP science. The resulting report, *Bringing the Genome to Life* (August 2000), set forth recommendations that led to the *Genomes to Life* roadmap (April 2001). Funding for FY 2002 was $21.7 million. The FY 2003 budget for the program proposed in the President's Request to Congress is $42.4 million.

U.S. Department of Energy
Office of Science

Marvin Frazier
Office of Biological and Environmental Research (SC-72)
301/903-5468, Fax: 301/903-8521
marvin.frazier@science.doe.gov

Gary Johnson
Office of Advanced Scientific Computing Research (SC-30)
301/903-5800, Fax: 301/903-7774
gary.johnson@science.doe.gov

Web site for this document:

- **DOEGenomesToLife.org/pubs/overview.pdf**

GTL Program Projects

Harvard Medical School

Microbial Ecology, Proteogenomics and Computational Optima

A2

Microbial Ecology, Proteogenomics, and Computational Optima

George Church (church@arep.med.harvard.edu), Sallie Chisholm, Martin Polz, Roberto Kolter, Fred Ausubel, Raju Kucherlapati, Steve Lory, Mike Laub, Robert Steen, Martin Steffen, Kyriacos Leptos, Matt Wright, Daniel Segre, Allegra Petti, Jake Jaffe, David Young, Eliana Drenkard, Debbie Lindell, Eric Zinser, and Andrew Tolonen

Harvard Medical School and Massachusetts Institute of Technology

Understanding microbial cells and communities requires system models, not just subsystems, but comprehensive, genome-wide analyses. Genotype + environment yields phenotype. New methods allow us to cost-effectively "overdetermine" each of these three components enabling studies of mechanism, optimality, and bioengineering. The key to this will be integration of measures of molecules per cell of RNA, proteins and metabolites.

Beyond concentrations, we need to image and model 4D structures of cells and of communities of cells. New technologies include single-molecule sequencing with polymerase colonies (polonies) to assess RNA and DNA states. New genetic selections allow phenotypes of genome-wide sets of mutations using a microarray read-out. New computational approaches include "expression coherence" for combinations of transcription elements and "Minimization of Metabolic Adjustment" (MoMA) to model proliferation of mutants. We are applying these methods to *Prochlorococcus*, responsible for a major fraction of the earth's microbial carbon fixation, *Caulobacter*, relevant to dilute scavenging and bioremediation as well as cell division, *Pseudomonas*, displaying a broad range of metabolic pathways including chemical/biological toxins and well-studied biofilms, and to other species in their communities including "uncultivated isolates."

For more complete descriptions & updates see http://arep.med.harvard.edu/DOEGTL.

Lawrence Berkeley National Laboratory

Rapid Deduction of Stress Response Pathways in Metal/ Radionuclide Reducing Bacteria

A4

Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria

Adam Arkin[2,3] (aparkin@lbl.gov), Alex Beliaev[8], Inna Dubchak[2], Matthew Fields[1], Terry Hazen[2], Jay Keasling[2], Martin Keller[4], Vincent Martin[2, 3], Frank Olken[2], Anup Singh[5], David Stahl[7], Dorothea Thompson[1], Judy Wall[6], and Jizhong Zhou[1]

[1]Oak Ridge National Laboratory; [2]Lawrence Berkeley National Laboratory; [3]University of California, Berkeley; [4]Diversa, Inc.; [5]Sandia National Laboratories; [6]University of Missouri, Columbia; [7]University of Washington, Seattle; and [8]Pacific Northwest National Laboratory

The focus of our research is the characterization of regulatory networks in microorganisms, and the creation of data-driven, validated mathematical models of stress response to conditions commonly found in U.S. Department of Energy (DOE) metal and radionuclide contaminated sites. We have created an integrated program of applied environmental microbiology, functional genomic measurement, and computational analysis and modeling that seeks to understand the basic biology involved in a microorganism's ability to survive in the relevant contaminated environments while reducing metals and radionuclides. Our main focus is *Desulfovibrio vulgaris* because of its metabolic versatility, its ability to reduce metals of interest to DOE, and its relatively easy culturability and molecular biology. However, because achieving our programmatic goals requires a comparative analysis of regulation among multiple bacteria in the environment, we are also studying *Shewanella oneidensis* and *Geobacter metallireducen*s, which follow different lifestyles than *Desulfovibrio*. Because a strong research community is already studying these former two microbes' behavior under the auspices of DOE's Microbial Cell program, we are coordinating with those teams to jumpstart the initial research of this program. Our overarching goal is to develop criteria for monitoring the integrity (health) and altering the trajectory of an environmental biological system (process control). To achieve this requires a more complete understanding of how the biological "units" comprising the system are organized, regulated, and linked in time and space (genes, genomes, cells, populations, communities, and ultimately, ecosystems). Key to these objectives is a more complete understanding of stress response systems and their environmental context.

During the first few months of this project, we have established our three research cores in Applied Environmental Microbiology (AEMC), Function Genomics (FGC), and Computation (CC). Each core has established a work plan with specific tasks. The tasks and more detailed accomplishments of each core will be presented in separate posters. A Web page (http://vimss.lbl.gov) was established immediately for communication to the public, scientific community and the project teams. As part of the web page, we have established bulletin boards for discussion and an interface with the project database (Biofiles). Investigators have uploaded protocols for sampling and analysis, and data of various types to the Biofiles database that the Computational Core is developing for the project. The CC has obtained sequences for all three bacteria and begun analysis. The initial annotations have been curated, operon, regulon and cis-regulatory sequence prediction have been made and the visualization tools are now being built. The FGC has acquired new instrumentation (eg., Mass spectrometers) and begun testing on *Shewanella* strains. Standard culture conditions for the *Desulfovibrio* strains have also been tested at all sites and preliminary proteomics data has been obtained. The AEMC has documented available data from the Field Research Center at Oak Ridge from various investigators and begun rigorous analysis of samples for sulfate reducers and in particular *Desulfovibrio* strains. The AEMC has also acquired anaerobic chambers and sediment samples from contaminated areas at the FRC and begun analysis of stressors to determine the most appropriate initial simulations and directions for the project. The initial focus is on pH, N, P, and O.

Oak Ridge National Laboratory

Genomes to Life Center for Molecular and Cellular Systems

A Research Program for Identification and Characterization of Protein Complexes

A6

Bioinformatics and Computing in the Genomes to Life Center for Molecular and Cellular Systems

D. A. Payne*[1] (debbie.payne@pnl.gov), E. S. Mendoza[1], G. A. Anderson[1], D. K. Gracio*[1], W. R. Cannon[1], T. P. Straatsma[1], H. J. Sofia[1], D. A. Dixon*[1], M. Shah[2], D. Xu[2], D. Schmoyer[2], S. Passovets[2], I. Vokler[2], J. Razumovskaya[2], T. Fridman[2], V. Olman[2], A. Gorin[2], E. Uberbacher[2], F. Larimer[2], and Y. Xu[2]

*Presenters

[1]Pacific Northwest National Laboratory; and [2]Oak Ridge National Laboratory

Scientists will generate large amounts of experimental and computational data at the ORNL/PNNL Genomes to Life (GTL) Center for Molecular and Cellular Systems. Data will be generated at several collaborating facilities and will need to be shared among the collaborators and, ultimately, with the wider research community. The processing, analysis, management, and storage of this data will require a flexible, robust, and scalable information system. As the GTL project ramps up, many of the data and sample tracking and analysis functions will need to be automated and integrated to keep up with the high-throughput processes. Since the start of the project, our bioinformatics work has been focusing on three areas: 1) laboratory information management system (LIMS) in support of the Center's data management and storage, 2) mass spectrometry proteomics analysis, and 3) bioinformatic analysis tools.

LIMS System

We have purchased a commercially available and proven LIMS system, Nautilus™ (from Thermo Lab Systems) to serve as the backbone for integrating data management and analysis. Nautilus, once configured, will provide comprehensive sample tracking from planning through experimentation, data analysis, reporting, and final archival or disposal. Nautilus will be interfaced with laboratory instruments and data analysis tools and services to enable automation and standardization of data processing. Data will be archived through integration with the Environmental Molecular Sciences Laboratory Northwest File System archive.

A key to the success of this project will be the ability for users to have ubiquitous, seamless access to LIMS data at both ORNL and PNNL. To accomplish this data sharing, a schema will be defined for components and workflow that are common to both facilities, and software will be written to access data from both instances of the LIMS system. Current activities include defining the overall system, defining the data management schema for the respective facilities at ORNL and PNNL, gathering requirements, and identifying common data structures.

Mass Spectrometry Proteomic Data Analysis

Before the GTL program started, PNNL developed the Proteomics Research Information System and Management (PRISM) system that stores, tracks pedigree of, and provides automated analyses of proteomic data. PRISM will be used both at PNNL and at ORNL for mass spectrometry data analysis. It is composed of distributed software components that operate cooperatively on several commercially available computer systems that communicate over standard network connections. PRISM collects data files directly from all mass spectrometers in the laboratory and manages storage and tracking of these data files as well as automates the processing into both intermediate results and final products.

PRISM will be installed at ORNL to provide a common proteomic data analysis capability. Additionally, a mass spectrometry data analysis pipeline for automated processing of large-scale mass spectrometry data of proteins and protein complexes has been designed and is in the early stages of implementation. The pipeline is designed to process data generated using both bottom-up and top-down approaches and to combine information derived from both approaches for identifying proteins and protein complexes. The pipeline

builds a data interpretation capability based on three existing mass spectrometry data analysis software: SEQUEST, MASCOT, and COMET. These tools have been evaluated with systematic comparison using experimental data. Through these analyses, computational techniques have been developed for assessing the reliability of these identification tools. For example, in the case of SEQUEST, a neural network and a statistics-based method has been developed for such reliability assessment. Such a capability can significantly remove the need of human involvement in large-scale MS data interpretation. New methods for de novo sequencing that can complement database search-based methods for protein identification are also under development.

Bioinformatic Analysis Tools

In the area of bioinformatics, our project is focused in many areas: computational inferencing of protein complexes, including membrane-associated complexes, dynamic simulation of protein-protein interaction, and functional mechanism studies of protein complexes; characterizations of amino acids and peptide transport pathways; and identification of operons and regulons. Interactive analysis and visualization tools are being developed to support these goals.

This research is 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.

A8

Mass Spectrometry in the Genomes to Life Center for Molecular and Cellular Systems

Gregory B. Hurst[1] (hurstgb@ornl.gov), Robert L. Hettich[1], Nathan C. Verberkmoes[1], Gary J. Van Berkel[1], Frank W. Larimer[1], Trish K. Lankford[1], Steven J. Kennel[1], Dale Pelletier[1], Jane Razumovskaya[1], Richard D. Smith[2], Mary Lipton[2], Michael Giddings[5], Ray Gesteland[4], Malin Young[3], and Carol Giometti[6]

[1]Oak Ridge National Laboratory; [2]Pacific Northwest National Laboratory; [3]Sandia National Laboratories; [4]University of Utah; [5]University of North Carolina; and [6]Argonne National Laboratory

Mass spectrometry is a significant contributor to the Center for Molecular and Cellular Systems due to its capability for high-throughput identification of proteins and, by extension, protein complexes. From the outset of the Genomes To Life (GTL) Program, therefore, mass spectrometry has an important role to play in the pursuit of Goal 1 of the GTL—the identification of the "machines of life." The potential utility of mass spectrometry to GTL, however, extends far beyond current capabilities. In addition to incorporation of state-of-the- art mass spectrometry as a resource, we have also included a mass spectrometry research component as part of the Center for Molecular and Cellular Systems. The aim of this research component is to improve on existing mass spectrometry tools for protein complex characterization, as well as to produce new tools that will further the goals of the GTL program. Key to the success of this research component is close interaction with the protein expression, complex isolation, computational and imaging components of the Center.

Currently, mass spectrometry is contributing heavily to the process of identifying target proteins that are likely to be members of complexes in *Rhodopseudomonas palustris*. These target proteins will be evaluated for expression as fusions with affinity labels to facilitate isolation of complexes. This identification process is based on mass spectrometric detection, in pelleted fractions, of proteins that one would normally expect to find in soluble fractions, indicating possible membrane association or membership in a large complex. From MS analysis of proteins from two different growth conditions of *R. palustris*, an initial list of target proteins has been assembled. The

MS analysis strategy at ORNL measures both intact molecular masses ("top-down") and tandem mass spectra of tryptic digests of proteins ("bottom-up"). The "bottom-up" approach allows more comprehensive identification of proteins in a sample, while the "top-down" approach, which exploits the high-performance characteristics of Fourier transform mass spectrometry, provides information on post-translational modifications. The accurate mass tag (AMT) approach at PNNL is aimed at increasing throughput, sensitivity, and dynamic range for enhancing the detection of low-copy-number proteins and complexes.

We have also obtained initial mass spectrometry results from affinity purifications of fusions of *R. palustris* genes with GST and 6-HIS affinity tags, expressed in *E. coli*, verifying correct expression of the fusion proteins. Two strategies are being compared for this measurement. The first strategy is to elute affinity-captured proteins from the resin, separate by 1D SDS-PAGE, excise bands, digest, and analyze by reverse-phase nanoscale liquid chromatography on line with nano-electrospray/tandem mass spectrometry. The second strategy is to eliminate the gel separation, and simply digest the entire mixture eluted from the affinity resin. The latter strategy will improve throughput considerably. "Top-down" measurements of affinity-captured fusion proteins are also underway. Current experiments directed toward expression of affinity-labeled proteins in *R. palustris* will provide our first opportunity for mass spectrometric identification of proteins that associate with these labeled targets-an important first step for Goal 1 of GTL.

Combined mass spectrometric and computational methods for characterizing crosslinked protein complexes are also under development. Crosslinking offers the opportunity to stabilize "fragile" complexes. Furthermore, it provides an alternative method to introduce an affinity tag into a protein complex, potentially increasing the throughput of analysis of complexes. Technical issues to be solved include increasing the robustness of crosslinking protocols, mass spectrometric detection of crosslinks, and computational methods for data interpretation. We have made progress in optimizing an affinity purification procedure based on peptides that have been crosslinked using a biotinylated reagent. Computer programs for interpretation of mass spectra of crosslinked samples have been initiated. Demonstration of integrating these various components on a model protein complex is underway.

Although not all funded by GTL, other mass spectrometric techniques relevant to the goals of the GTL are also under development. At ORNL, these include a method for characterizing surfaces of proteins and protein complexes via oxidative chemistry combined with mass spectometry, and sampling by electrospray mass spectrometry of proteins captured on surfaces displaying arrays of affinity-capture reagents surfaces. PNNL is developing hardware improvements for increasing the speed, sensitivity, and dynamic range of measurements, as well as informatic methods for incorporating chromatography elution information in protein identification techniques.

This research sponsored by Office of Biological and Environmental Research, U.S. Department of Energy. Oak Ridge National Laboratory (ORNL) is managed by UT-Battelle, LLC, for the U. S. Department of Energy under Contract No. DE-AC05-00OR22725.

A10

Genomes to Life Center for Molecular and Cellular Systems: A Research Program for Identification and Characterization of Protein Complexes

Joshua N. Adkins[1], Deanna Auberry[1], Baowei Chen[1], James R. Coleman[1], Priscilla A. Garza[1], Jane M. Weaver Feldhaus[1], Michael J. Feldhaus[1], Yuri A. Gorby[1], Eric A. Hill[1], Brian S. Hooker[1], Chian-Tso Lin[1], Mary S. Lipton[1], L. Meng Markillie[1], M. Uljana Mayer[1], Keith D. Miller[1], Sewite Negash[1], Margaret F. Romine[1], Liang Shi[1], Robert W. Siegel[1], Richard D. Smith[1], David L. Springer[1], Thomas C. Squier[1], **H. Steven Wiley**[1] (steven.wiley@pnl.gov), Linda J. Foote[2], Trish K. Lankford[2], Frank W. Larimer[2], T-Y. S. Lu[2], Dale Pelletier[2], Stephen J. Kennel[2], and Yisong Wang[2]

[1]Pacific Northwest National Laboratory; and [2]Oak Ridge National Laboratory

Summary: We have developed methodologies for isolating and identifying multiprotein complexes in *Shewanella oneidensis* MR-1 (PNNL) and *Rhodopseudomonas palustris* (ORNL), whose metabolisms are important in both understanding microbial energy production and environmental remediation. We are comparing complementary methods involving the isolation and identification of transient and stable protein complexes, with a current focus on validating the physiological relevance of isolated protein complexes.

Cloning, Expression, and Purification: To date, 23 *S. oneidensis* genes have been cloned into the GATEWAY™ expression vector pDEST™ containing a His_6-tag for purification. Initial screening tests indicate that ~73% of cloned genes were expressed. Among those expressed proteins, 8 were purified to homogeneity using a Ni-NTA column under nondenaturing conditions. The yields of purified proteins obtained from 1 L of culture varied from 5 to 29 mg. We have also constructed new GATEWAY™-compatible vectors that will permit the expression of His_6-tagged proteins in both *S. oneidensis* and *R. palustris* and the subsequent isolation of preformed complexes from microbes. Using four modified pDEST vectors, 7 *R. palustris,* genes have been cloned and expressed in *E. coli*. We are testing both N and C-terminal 6-his and GST tags for efficiency of expression and purification . Western blots of proteins and MS spectra of tryptic digests (see MS poster) of the GST-tagged nitrite reductase verify the expression and purification of polyproteins at high yield. The modified vector containing the GroEL gene has been inserted into *R. palustris* and it appears to be retained and convey drug resistance to the bacteria. Pull down experiments are in progress to isolate complexes from this target organism.

Affinity Reagent Generation: Purified proteins from *S. oneidensis* are currently being screened against a cell surface display of single-chain fragment variable (scFv) antibodies on the yeast *Saccharomyces cerevisiae* developed at PNNL, allowing rapid generation of affinity reagents that will permit the capture of protein complexes formed *in vivo*. We expect that these affinity reagents will cross-react with homologous protein complexes in different microbes, permitting the rapid isolation of protein complexes in a generalized manner.

Tagging and Cross-Linking Approaches for Complex Isolation: In addition to the His6-tag, additional epitope tags are being assessed for their utility in enhancing the specificity of complex isolation under milder isolation conditions that will retain low-affinity binding partners in protein complexes. To date, we have demonstrated the utility of the CCXXCC epitope sequence for protein purification. Likewise, commercially available light-activated cross-linking reagents have been used to stabilize protein complexes in cellular homogenates from *Shewanella*, permitting the affinity purification of protein complexes under more stringent conditions that remove nonspecifically associated proteins. Under these conditions a limited range of cross-linked products are observed that are readily characterized by mass spectrometry.

Complex Isolation and Identification: Critical to the development of robust methods to rapidly isolate protein complexes is the assessment of standard protocols to isolate and identify different classes of protein complexes. We have therefore developed parallel methods focusing on the isolation and identification of membrane and soluble protein complexes that are known to form either stable or transient protein-protein interactions. Initial measurements have focused on the identification of stable and soluble protein complexes (e.g., RNA polymerase A), which has permitted the validation of protein isolation and cross-linking methods and the development of conditions that minimize nonspecific protein associations. However, because dynamic changes in protein complexes are expected to provide important insights into the metabolic regulatory strategies used by these organisms to adapt to environmental changes, we have extended these methods to assess transient protein interactions associated with signal transduction proteins (phosphotyrosine phosphatase A) and stress-regulated proteins (e.g., methionine sulfoxide reductases A and B). In the latter cases, these proteins are known to interact and reduce oxidized substrates on a time scale of minutes. The development of immunoprecipitation methods that permit the isolation of transient complexes involving these proteins suggests that generalizable strategies to rapidly isolate protein complexes can be used to identify the formation of transient protein complexes. Surprisingly, the catalytic activity of methionine sulfoxide reductases from *Shewanella* has additional catalytic activities relative to those found in either *E. coli* or vertebrates, consistent with *Shewanella*'s known ability to thrive under harsh environmental conditions. We expect that identifying binding partners between this critical antioxidant protein will, furthermore, provide important information regarding oxidatively sensitive proteins and associated regulatory strategies that these organisms implement to survive.

Of the 7 *R. palustris* proteins expressed in the modified pDEST vector, we are concentrating on the GroEL chaparonin protein to validate complex formation. The tagged protein expressed in *E. coli* can be used to complex with GroES from *R. palustris* to document complex formation and pull-down efficiency. *R. palustris* has two different genes for GroEL type proteins and we will test if

each is expressed and if they form co-complexes or if they are used separately for different functions. Dissimilatory nitrite reductases are capable of generating a membrane potential, as well as providing an electron sink for maintenance of balanced photosynthetic growth in the presence of highly reduced C-sources. In addition, there is a report that cells engaged in denitrification have an altered chemotactic response. Other systems being expressed include subunits of the uptake hydrogenase and components of sulfite oxidation, i.e., sulfite dehydrogenase, and sulfite oxidase.

This research is 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. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U. S. Department of Energy under Contract No. DE-AC05-00OR22725.

A12

New Approaches for High-Throughput Identification and Characterization of Protein Complexes

Michelle Buchanan[1] (buchananmv@ornl.gov), Frank Larimer[1], Steven Wiley[2], Steven Kennel[1], Thomas Squier[2], Michael Ramsey[1], Karin Rodland[2], Gregory Hurst[1], Richard Smith[2], Ying Xu[1], David Dixon[2], Mitchel Doktycz[1], Steve Colson[2], Carol Giometti[3], Raymond Gesteland[4], Malin Young[5], and Michael Giddings[6]

[1]Oak Ridge National Laboratory; [2]Pacific Northwest National Laboratory; [3]Argonne National Laboratory; [4]University of Utah; [5]Sandia National Laboratories; and [6]University of North Carolina

The Center for Molecular and Cellular Systems (CMCS) is a recently established project that focuses specifically on Goal 1 of the GTL program. Its aim is to identify and characterize the complete set of protein complexes within a cell to provide a mechanistic basis of biochemical functions. Achieving this Goal would provide the ability to understand cells and their components in sufficient detail to allow the creation of network maps of cells that could be used in building models to predict, test and understand the responses of a biological system to its environment. Further, Goal 1 forms the foundation necessary to accomplish all of the other objectives of the GTL program, which are focused on gene regulatory networks and molecular level characterization of interactions in microbial communities.

A stated goal of the GTL program is to identify greater than 80% of the protein complexes in an organism per year within the first five years of the program. Ultimately, the GTL program will require the analysis of thousands of protein complexes from hundreds of microbes each year. The central task of the CMCS (Core Project) is to integrate biological, analytical, and computational tools to allow identification and characterization of protein complexes in a robust, high-throughput manner. The Core includes systems for growth of microbial cells under well-characterized conditions, isolation of protein complexes from cells, and their analysis by mass spectrometry (MS), followed by verification and characterization by imaging techniques. Several approaches for the isolation of the complexes are currently being examined and compared, including affinity tags (e.g., GST and 6-HIS affinity tags) and single chain antibodies. Computational tools are being integrated into this process to track samples, interpret the data, and to archive and disseminate data. Automated, parallel sample handling processes will be incorporated to maximize throughput and minimize amount of sample required.

The CMCS is initially focused on the identification and characterization of protein complexes in two microbial systems, *Shewanella oneidensis* and *Rhodopseudomonas palustrus*. The aim is to obtain a knowledge base that can provide insight into the relationship between the complement of protein complexes in these microbes and their biological function. Early activities within the Core have focused on setting up isolation, purification and analysis techniques and obtaining data on specific complexes in these two microbes. For *R. palustris*, we have performed baseline growth studies in two important metabolic states, anaerobic photohetero-trophic and dark aerobic heterotrophic. Wild-type cultivations at up to 2-L have generated samples for proteome analysis and for isolation of protein complexes. Data has been obtained from affinity purification of fusion proteins between several *R. palustris* genes and GST and 6-HIS affinity tags have been expressed in *E. coli*. We have verified correct expression of the fusion proteins and affinity-labeled proteins in *R. palustrus*. Various forms of chaperonin60, nitrite reductase, hydrogenase subunits, sulfite dehydrogenase, and thiosulfite oxidase are currently being examined. Work with *Shewanella* has focused on an initial set of tagged proteins expressed in *E. coli*; 20 proteins are in progress,

among them, phosphotyrosine phosphatase, methionine sulfoxide reductase and RNA polymerase–alpha subunit have been purified and carried forward to use as bait with *Shewanella* extracts, with MS-MS analysis proceeding.

The Core of the CMCS will generate large amounts of experimental data at different sites and these data will need to be shared among the collaborators and, ultimately, with the wider research community. The management and storage of this data requires a flexible, robust and scalable information system. After a comprehensive analysis and evaluation of the CMCS's process and data flow information need, we selected a Laboratory Information Systems (LIMS) that will serve as the backbone for integrating data management and analysis. Concurrent with evaluation of LIMS systems, we have also examined the processes within the Core that can be readily automated and incorporated into parallel processes (e.g., 96 well plate format), such as cell lysis, complex isolation, and final purification prior to MS analysis.

As initial data are generated within the Core, we are also evaluating the technologies to identify bottlenecks and needs for technology improvement. Current technologies for the identification and characterization of protein complexes will not be sufficient to meet the long-term goals of the GTL program. Therefore, a number of research tasks have been devised to address specific requirements of the Core, including new approaches for high throughput complex processing. For example, as part of the efforts to improve sample processing, we are evaluating microfluidic devices for microbial cell lysis and protein/peptide separation. We are also examining novel approaches for optimizing molecular characterization by MS, such as improving sensitivity and dynamic range. Combined MS and computational methods for characterizing crosslinked protein complexes area also under development. Crosslinking offers the opportunity to stabilize "fragile" complexes, and is an alternative to introducing an affinity tag into the complex, potentially increasing analysis throughput. Initial investigations have included optimization of an affinity purification procedure based on crosslinked biotinylated peptides, and the identification of putative cross-links in model protein complexes. In addition, imaging techniques are being developed to validate the presence of complexes in cells and to provide physical characterization of the complexes. Finally, bioinformatics tools for data tracking, acquisition, interpretation, and dissemination, along with computational tools for modeling and simulation of protein complexes are being developed.

This research sponsored by Office of Biological and Environmental Research, U.S. Department of Energy. Oak Ridge National Laboratory (ORNL) is managed by UT-Battelle, LLC, for the U. S. Department of Energy under Contract No. DE-AC05-00OR22725.

A14

Automation of Protein Complex Analyses in *Rhodopseudomonas palustris* and *Shewanella oneidensis*

P. R. Hoyt[1] (hoytpr@ornl.gov), C. J. Bruckner-Lea[2], S. J. Kennel[1], P. K. Lankford[1], M. S. Lipton[2], R. S. Foote[1], J. M. Ramsey[1], K. D. Rodland[2], and M. J. Doktycz[1]

[1]Oak Ridge National Laboratory; and [2]Pacific Northwest National Laboratory

High-throughput analyses afforded by mass spectroscopy require sample preparation processes that can keep pace. Standardization and automation of protein "pulldowns", and related reagents are being developed. The processes are designed to provide a straightforward material flow in high-throughput format for the pulldown of protein complexes from the *Rhodopseudomonas palustris* and *Shewanella oneidensis* genomes. Existing techniques are well developed; however, some processes in clone library, antibody, and protein complex production have never been automated and few established protocols are available. In order to provide the highest level of biological significance and protein interaction coverage, the protein complex pulldowns from the different organisms will use different strategies. Subsequently, automation is designed to use flexible, compatible processes of varied scale during the program such that advances in technology can be evolved into innovative high-throughput techniques for sample preparation. The result will be a unique and robust system for protein expression and complex pulldown in bacterial systems.

The process for production of native tagged proteins for complex pulldown experiments uses conventional fluidics scale of 96-well format and liquid handling robotics. It is subdivided into the molecular preparation of a complete genomic library of expression clones for in vivo expression of *R. palustris* genes, followed by the production

of proteins and "pull-downs" of protein complexes for analyses by mass spectrometry. The gene library and protein production scheme involves a suite of high-throughput molecular biology techniques based on the Gateway™ technology cloning strategy supplied by Invitrogen Corporation. This process requires two rounds of recombination between purified DNAs to produce protein expression vectors suitable for pull-down experiments in RP. At this time, all PCR setup, PCR purification, plasmid isolation, and redistribution steps, have been fully automated and integrated into an information management system for sample tracking. Recombination reactions should be fully automated in the near future using existing instrumentation. High-throughput automation of the electroporation steps, as well as colony picking can be automated using commercially available products, which are currently being evaluated. This leaves only the plating of bacteria on selective media to rely on manual processes.

Because detergents are not compatible with mass spectroscopic analyses, manual disruption processes were required. We were able to adapt a high-throughput, closed container non-detergent bead-milling technology (used originally for high-throughput isolation of RNA from animal tissues), to disrupt the *R. paulutris* cell walls. This process results in comparable protein profiles generated using other physical disruption techniques. Bead milling has been found to be most compatible with downstream MS analyses. Additionally, it reduces cross-contamination, and provides an extraordinary level of automation to the production process.

An heterologous-tagged protein pulldown system, for *S. oneidensis* using single-chain antibodies (Ab) to specific expressed proteins is also under automation development. This process uses a microfluidics platform combined with functionalized microbeads for the purification of protein complexes. A renewable microcolumn system with optical detection has been assembled and automated procedures developed. The renewable microcolumn consists of small volumes (microliters) of microbeads that are automatically packed, perfused with cell lysates, and wash solutions, and proteins eluted using a solution that is suitable for mass spectrometry analysis. After each purification, the small volume of microbeads is automatically flushed from the microcolumn and a new microcolumn is automatically packed. The microbeads are functionalized for the capture of a specific protein, for example by derivatization with an antibody for the protein of interest. Optical monitoring of the microcolumn during processing provides information about the amount of material on the column during each binding and washing step. The current automated procedure can process a cell lysate volume ranging from 10 microliters to 1 millilter, and the purified proteins are eluted into 150 microliters of a low salt buffer solution. Automated procedures are currently being tested for the capture of *Shewanella* proteins tagged with yellow fluorescent protein (YFP), along with the proteins that associate with the YPF-tagged protein. As new reagents for protein capture such as single chain antibodies for *Shewanella* proteins of interest are developed, they will be linked to microbeads and renewable column protocols will be developed for automated purification of the protein complexes for mass spectrometry. In the next stage of this work, the eluted protein complexes will be analyzed by mass spectrometry and the automated protocols will be optimized.

For the ultimate in throughput and sensitivity, a lab-on-a-chip complex isolation and identification program is also under development. Many of the individual steps involved in sample processing and analysis, including cell lysis, protein/peptide separations and enzyme digestions, have been implemented in microfluidic devices that can be interfaced with mass spectrometry for on-line analysis. (We have previously demonstrated electrically induced lysis of mammalian cells in microfluidic devices and will apply this technique to bacterial protoplasts). The integration of these functions with a pull-down step would provide high-throughput analyses of protein complexes in extremely small numbers of cells.

In summary, protein complex analysis by mass spectroscopy will require a high-throughput reagent production scheme. Because the complexes isolated are different for the different organisms, different schemes for complex isolation have been implemented. At scales ranging from macro to micro we are automating the production of reagents and samples to produce these different complexes, and the processes are being optimized to feed into mass spectroscopic analyses. The automation development is concomitant with establishment of sample tracking and information management processes so that integration of these systems will be seamless.

This research sponsored by Office of Biological and Environmental Research, U.S. Department of Energy. Oak Ridge National Laboratory (ORNL) is managed by UT-Battelle, LLC,

for the U. S. Department of Energy under Contract No. DE-AC05-00OR22725.

Sandia National Laboratories

Carbon Sequestration in *Synechococcus*

From Molecular Machines to Hierarchical Modeling

A16

Analysis of Protein Complexes from a Fundamental Understanding of Protein Binding Domains and Protein-Protein Interactions in *Synechococcus* WH8102

Anthony Martino[1] (martino@sandia.gov), Andrey Gorin[2], Todd Lane[1], Steven Plimpton[1], Nagiza Samatova[2], Ying Xu[2], Hashim Al-Hashimi[3], Charlie Strauss[4], Byung-Hoon Park[2], George Ostrouchov[2], Al Geist[2], William Hart[2], and Diana Roe[1]

[1]Sandia National Laboratories, P.O. Box 969, MS9951, Livermore, CA 94551; [2]Oak Ridge National Laboratory, P.O. Box 2008, MS6367, Oak Ridge, TN 37831; [3]University of Michigan, Department of Chemistry, 930 N. University, Ann Arbor, MI 48109; and [4]Los Alamos National Laboratories, P.O. Box 1663, Los Alamos, NM 87545

The goal of this work is to characterize protein complexes in *Synechococcus* WH8102 by studying protein-protein interaction domains. We are focused on two efforts, one on the protein composition and cognate binding partners in the carboxysome, and another on characterization of known protein binding domains throughout the genome. An experimental design is chosen to integrate a number of computational techniques in order to develop a fundamental understanding of how protein complexes form.

Experimental Elucidation of Protein Complexes

Initial efforts will focus on the carboxysome, a polyhedral inclusion body that consists of a protein shell surrounding ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO). While RuBisCO regulates photosynthetic carbon reduction, the function of the carboxysome is unclear. The carboxysome may either actively promote carbon fixation by concentrating CO_2 or passively play a role by regulating RuBisCO turnover. The presence of carbonic anhydrase, an enzyme that regulates the equilibrium between inorganic carbon species, in the carboxysome would suggest an active role in carbon concentration, but experimental results are mixed. No clear biochemical evidence for a link between carbonic anhydrase and the carboxysome exists in WH8102.

We are developing synergistic techniques including protein identification mass spectrometry, yeast 2-hybrid, phage display, and NMR to characterize the composition, cognate binding partners, and protein interaction domains in the carboxysome. Established techniques are in progress to purify carboxysomes. Earlier literature indicate the carboxysome is composed of 5-15 peptides. In several organisms, a number of proteins within carboxysomes are known, and in *Synechococcus* WH8102, a number are inferred by homology. Results are dependent on sometimes difficult carboxysome preparations. We hope to report on progress in this area specific to *Synechococcus* WH8102. After SDS-PAGE separation and in-gel enzymatic digests, comprehensive protein identification will be determined using quadrupole time-of-flight mass spectrometry with an electrospray ionization source. Cognate binding pairs between known proteins will be determined by systematic yeast 2-hybrid experiments. The results will be verified and explored further using phage display to determine potential protein binding domains. Both genomic and random peptide libraries will be employed. Finally, we will pursue the development of automated RDC-NMR methods for high throughput assignments and characterization of relative domain alignments in two sub-units in RuBisCO (52 KDa) and organization of the carboxysome. NMR methods for characterizing protein-protein interactions are also being developed that rely on probing interactions between proteins and peptide moieties that are attached to field oriented phage particles. Such an approach would enjoy high sensitivity to molecular interactions, providing an effective complement to phage display methods.

In a broader effort, proteins in *Synechococcus* WH8102 containing known binding domains will be explored using phage display. Eight TPR, four PDZ, and four CBS domains are indicated by pfam analysis in ORFs of *Synechococcus* WH8102. Three SH3-homologous domains have been described in other cyanobacteria. Determination of consensus binding sites within the genome will characterize possible fundamental interaction domains in complexes and provide insight for computing theoretical protein interaction maps.

Computational Elucidation of Protein Complexes

Investigations of protein-protein interactions are conducted on many levels and with different questions in mind–ranging from the reconstruction of genome-wide protein-protein interaction networks and to detailed studies of the geometry/affinity in a particular complex. Yet as the questions asked at the different levels are often intricately related and interconnected, we are approaching the problem from several directions, developing computational methods involving sequence analysis approaches, low resolution prediction of protein folds and detailed atom-atom simulations.

The sequencing of complete genomes has created unique opportunities to *fuse the knowledge* extracted from genomic contexts for prediction of the functional interactions between genes. Here we demonstrate that unusual protein-profile pairs can be "learned" from the database of experimentally determined interacting proteins. Distributions of protein-profile counts are calculated for random and interacting protein pairs. A pair of protein-profiles is considered unusual if its frequency distribution is significantly different compared to what is expected at random. We demonstrate that statistically significant patterns can be identified among protein-profiles characterized by the PFAM domains, Blocks protein families, or InterPro signatures but not by the PROSITE and TIGRFAM. Such patterns can be used for predicting putative pairs of interacting proteins beyond original "learning database".

In addition to "sequence-based" protein signatures one of our main aims is the development of structure-based algorithms for the inference of protein-protein interactions. At the initial stage we will apply structure prediction methods to determine protein fold families with our ROSETTA and PROSPECT programs and use inferred structural similarities to create hypotheses about their interacting partners. The necessary step in this process is a creation of structure prediction pipeline for high throughput characterization of the protein folds. The computational pipeline merges several bioinformatics and modeling tools including algorithms for protein domain division, secondary structure prediction, fragment library assembly, and structure comparison. Since the protein folding algorithms deliver not unique answers but rather ensembles of predictions we will also construct database system to store and curate the accumulated inferences.

Finally, we are developing tools for full atom modeling of protein-protein interactions. Tempering capability is being integrated into our parallel molecular dynamics code (LAMMPS). In tempering, multiple copies of a system are simulated simultaneously. Temperature exchanges are performed between copies to more efficiently sample conformational space. We are using tempering to generate conformations of short peptide chains in solution, similar to the peptide fragments that bind to proteins in the phage display experiments our team is performing. These conformations will be used in peptide docking calculations against protein binding domains from *Synechococcus*. We are extending our docking code PDOCK with genetic-algorithm optimizers to enable peptide flexibility in this step. The computed conformations of docked complexes will be further relaxed and solvated with molecular tools (MD and classical DFT) to estimate relative binding affinities, converting experimental phage display output into quantitative protein/protein network data.

A18

Carbon Sequestration in *Synechococcus*: Microarray Approaches

Brian Palenik[4], Anthony Martino[2], **Jerilyn A. Timlin**[2] (jatimli@sandia.gov), David M. Haaland[2], Michael B. Sinclair[2], Edward V. Thomas[2], Vijaya Natarajan[3], Arie Shoshani[3], Ying Xu[1], Dong Xu[1], Phuongan Dam[1], Bianca Brahamsha[4], Eric Allen[4], and Ian Paulsen[5]

[1]Oak Ridge National Laboratory; [2]Sandia National Laboratories; [3]Lawrence Berkeley National Laboratory; [4]Scripps Institute of Oceanography; University of Southern California, San Diego; and [5]The Institute for Genomic Research

Synechococcus sp. are major primary producers in the marine environment. Their carbon fixation rates are likely affected by physical and chemical factors such as temperature, light, and the availability of nutrients such as nitrate and phosphate. In our GTL, microarray analysis is being developed as a collaborative multidisciplinary project to characterize *Synechococcus* gene expression under different environmental stresses. We are constructing a whole genome microarray. We are developing microarray experiments using statistical considerations as input to the process. We are analyzing the arrays with a unique hyperspectral scanner and associated analysis algorithms. The microarray data will be archived using state of the art database management techniques. The microarray data will then be analyzed using our recently developed techniques for cluster, data mining, and incorporated in pathway analyses. The result will be biological insights into *Synechococcus* and marine primary productivity not achievable by a single investigator approach.

A20

Carbon Sequestration in *Synechococcus sp*.: From Molecular Machines to Hierarchical Modeling

Grant S. Heffelfinger[1](gsheffe@sandia.gov), Anthony Martino[2], Andrey Gorin[3], Ying Xu[3], Mark D. Rintoul III[1], Al Geist[3], Hashim M. Al-Hashimi[8], George S. Davidson[1], Jean Loup Faulon[1], Laurie J. Frink[1], David M. Haaland[1], William E. Hart[1], Erik Jakobsson[7], Todd Lane[2], Ming Li[9], Phil Locascio[2], Frank Olken[4], Victor Olman[2], Brian Palenik[6], Steven J. Plimpton[1], Diana C. Roe[2], Nagiza F. Samatova[3], Manesh Shah[2], Arie Shoshani[4], Charlie E. M. Strauss[5], Edward V. Thomas[1], Jerilyn A. Timlin[1], and Dong Xu[2]

[1]Sandia National Laboratories, Albuquerque, NM; [2]Sandia National Laboratories, Livermore, CA; [3]Oak Ridge National Laboratory, Oak Ridge, TN; [4]Lawrence Berkeley National Laboratory, Berkeley, CA; [5]Los Alamos National Laboratory, Los Alamos, NM; [6]University of California, San Diego; [7]University of Illinois, Urbana/Champaign; [8]University of Michigan; and [9]University of California, Santa Barbara

This talk will discuss the Sandia-led Genomes to Life (GTL) project: "Carbon Sequestration in *Synechococcus sp.*: From Molecular Machines to Hierarchical Modeling." This project is focused on developing, prototyping, and applying new computational tools and methods to ellucidate the biochemical mechanisms of the carbon sequestration of *Synechococcus sp.*, an abundant marine cyanobacteria known to play an important role in the global carbon cycle. Our effort includes five subprojects: an experimental investigation, three computational biology efforts, and a fifth which deals with addressing computational infrastructure challenges of relevance to this project and the Genomes to Life program as a whole. Some detail will be provided in this talk about each of our subprojects, starting with our experimental effort which is designed to provide biology and data to drive the computational efforts and includes significant investment in developing new experimental methods for uncovering protein partners, characterizing protein complexes, identifying new binding domains. Discussion of our computational efforts will include coupling molecular simulation methods with knowledge discovery from diverse biological data sets for high-throughput discovery and characterization of protein-protein complexes and developing a set of novel capabilities for inference of regulatory pathways in microbial genomes across multiple sources of information through the integration of computa-

tional and experimental technologies. We are also investigating methods for combining experimental and computational results with visualization and natural language tools to accelerate discovery of regulatory pathways and developing set of computational tools for capturing the carbon fixation behavior of complex of *Synechococcus* at different levels of resolution. Finally, because the explosion of data being produced by high-throughput experiments requires data analysis and models which are more computationally complex, more heterogeneous, and require coupling to ever increasing amounts of experimentally obtained data in varying formats, we have also established a companion computational infrastructure to support this effort. This element of our project will be discussed in the larger GTL program context as well.

A22

Systems Biology Models for *Synechococcus sp.*

Mark D. Rintoul[1] (rintoul@sandia.gov), Damian Gessler[2], Jean-Loup Faulon[1], Shawn Means[1], Steve Plimpton[1], Tony Martino[2], and Ying Xu[3]

[1]Sandia National Laboratories; [2]National Center for Genome Resources; and [3]Oak Ridge National Laboratory

Ultimately, all of the data that is generated from experiment must be interpreted in the context of a model system. Individual measurements can be related to a very specific pathway within a cell, but the real goal is a systems understanding of the cell. Given the complexity and volume of experimental data as well as the physical and chemical models that can be brought to bear on subcellular processes, systems biology or cell models hold the best hope for relating a large and varied number of measurements to explain and predict cellular response. Clearly, cells fit the working scientific definition of a complex system: a system where a number of simple parts combine to form a larger system whose behavior is much harder to understand. The primary goal of this subproject is to integrate the genomic data generated from the overall project's experiments and lower level simulations, along with data from the existing body of literature, into a whole cell model that captures the interactions between all of the individual parts. It is important to note here that all of the information that is obtained from other efforts in this project is vital to the work here. In a sense, this is the "Life" of the "Genomes to Life" theme of this project.

The precise mechanism of carbon sequestration in *Synechococcus sp*. is poorly understood. There is much unknown about the complicated pathway by which inorganic carbon is transferred into the cytoplasm and then converted to organic carbon. While work has been carried out on many of the individual steps of this process, the finer points are lacking, as is an understanding of the relationships between the different steps and processes. Understanding the response of *Synechococcus sp.* to different levels of CO_2 in the atmosphere will require a detailed understanding of how the carbon concentrating mechanisms in *Synechococcus sp.* work together. This will require looking these pathways as a system.

The aims of this part of the project are to develop and apply a set of tools for capturing the behavior of complex systems at different levels of resolution for the carbon fixation behavior of *Synechococcus sp*. The first aim is focused on protein network inference and deals with the mathematical problems associated with the reconstruction of potential protein-protein interaction networks from experimental work such as phage display experiments and simulation results such as protein-ligand binding affinities. Once these networks have been constructed, Aim 2 and Aim 3 describe how the dynamics can be simulated using either discrete component simulation (for the case of a manageably small number of objects) or continuum simulation (for the case where the concentration of a species is a more relevant measure than the actual number). Finally, in the fourth aim we present a comprehensive hierarchical systems model that is capable of tying results from many length and time scales together, ranging from gene mutation and expression to metabolic pathways and external environmental response.

University of Massachusetts, Amherst

Analysis of the Genetic Potential and Gene Expression of Microbial Communities Involved in the in situ Bioremediation of Uranium and Harvesting Electrical Energy from Organic Matter

A24

Analysis of the Genetic Potential and Gene Expression of Microbial Communities Involved in the in situ Bioremediation of Uranium and Harvesting Electrical Energy from Organic Matter

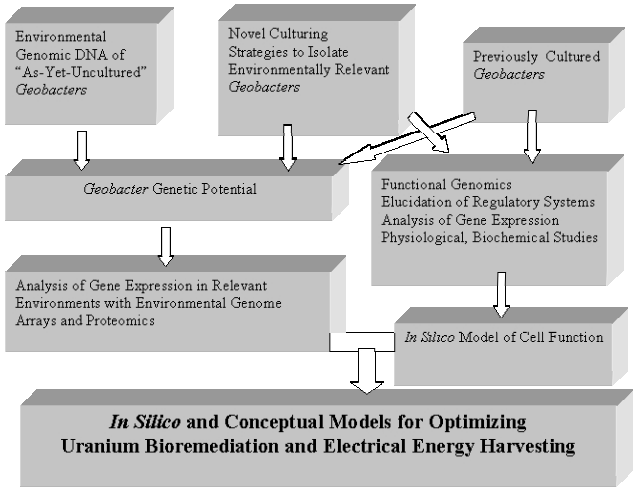
Derek Lovley[1] (dlovley@microbio.umass.edu), Stacy Ciufo[1], Zhenya Shebolina[1], Abraham Esteve-Nunez[1], Cinthia Nunez[1], Richard Glaven[1], Regina Tarallo[1], Daniel Bond[1], Maddalena Coppi[1], Pablo Pomposiello[1], Steve Sandler[1], Barbara Methé[2], Carol Giometti[3], and Julia Krushkal[4]

[1]University of Massachusetts; [2]The Institute for Genomic Research; [3]Argonne National Laboratory; and [4]University of Tennessee

The goal of this research is to develop models that can describe the functioning of the microbial communities involved in the in situ bioremediation of uranium-contaminated groundwater and harvesting electricity from waste organic matter. Previous studies have demonstrated that the microbial communities involved in uranium bioremediation and energy harvesting are both dominated by microorganisms in the family *Geobacteraceae* and that these *Geobacteraceae* are responsible for the uranium bioremediation and electron transfer to electrodes. The research plan is diagrammed below.

Progress to Date: Although the physiology of pure cultures of *Geobacters* are being studied and modeled in detail, the degree of similarity in the genetic potential of the *Geobacters* in culture and those that predominate during uranium bioremediation or electrical energy harvesting is unknown. The environmental component of the studies in the first four months of this project have focused a NABIR-program site, located in Rifle, Colorado in which the addition of acetate to the subsurface stimulated the growth of *Geobacter* species and the removal of uranium from the groundwater. In order to evaluate the genetic potential of the *Geobacter* species involved in uranium bioremediation, which at times accounted for over 80% of the total microbial community in the groundwater, genomic DNA was extracted from sediments undergoing active uranium reduction and is now being sequenced at the Joint Genome Institute. Some of this data should be available by the time of the meeting and novel methods for assembling complete or nearly complete *Geobacter* genomes from this environmental genomic DNA will be presented. An additional strategy to learning more about the genetic potential of *Geobacters* living in the subsurface was to isolate the predominant *Geobacters* and sequence their genomes. Using a novel technique, we were able to isolate a *Geobacter* from the study site whose 16S rDNA sequenced matched a 16S rDNA sequence that was prevalent in clone libraries from the uranium reduction zone at the study site. The genome of this organism will be studied in detail in the next year.

In order for information on the genetic potential of *Geobacters* to be useful in predicting the activity of *Geobacters* during bioremediation or energy harvesting, it is important to understand how gene expression is regulated. Although *Geobacters* have previously been considered to be metabolically simple organisms with little regulation, sequencing the genomes of several *Geobacters* has



revealed that they have multiple complex regulatory systems. Therefore, a major goal of this project is to investigate regulatory mechanisms in *Geobacters*. For example, analysis of the *G. sulfurreducens* genome revealed that it is highly attuned to its environment with the largest number of signal transduction proteins of any fully sequenced bacterium. Investigation of these regulatory systems as well as other fur-like, fnr-like, and sigma factor systems are currently underway. A novel regulatory system, discovered in our Genomes-to-Life research, in which Fe(III) serves as a repressor signal controlling the expression of the fumarate reductase genes will also be described.

Details on other key components of this project which include: additional environmental studies on energy-harvesting electrodes; functional analysis of genomes of multiple species in the family *Geobacteraceae*; and gene expression and proteomics studies to be conducted on sediments will also be presented.

GTL Communication

B63

Communicating Genomes to Life

Anne E. Adamson, Jennifer L. Bownas, **Denise K. Casey**, Sherry A. Estes, Sheryl A. Martin, Marissa D. Mills, Kim Nylander, Judy M. Wyrick, Laura N. Yust, and **Betty K. Mansfield** (mansfieldbk@ornl.gov)

Life Sciences Division, Oak Ridge National Laboratory, 1060 Commerce Park, MS 6480; Oak Ridge, TN 37830

For the past 14 years, the Human Genome Management Information System (HGMIS) has focused on presenting Human Genome Project information and imparting knowledge to a wide variety of audiences. Our goal has been to help ensure that scientists could participate in and reap the scientific bounty of this revolution, that new generations of students could be trained in the science, and the public could make informed decisions regarding complicated genetics issues. Building on that experience, for the past 2 years HGMIS also has been involved in communicating about the DOE Office of Science Genomes to Life program, sponsored jointly by the Office of Biological and Environmental Research (BER) and the Office of Advanced Scientific Computing Research (OASCR).

The Genomes to Life systems biology program is a departure into a new territory of complexity and opportunity requiring contributions from teams of interdisciplinary scientists from the life, physical, and computing sciences, necessitating an unprecedented integrative approach to both the science and to science communication strategies. Because each discipline has its own perspective and language, effective communication, in addition to technical achievement, is highly critical to GTL's overall scientific coordination and success. Part of the challenge is to help groups speak the same language from the team-building and strategy-development phases through program implementation and the reporting of results to scientific and public audiences. Our mission is to inform and foster participation by the greater scientific community, science administrators, educators, students, and the general public. Specifically, GTL communications goals include the following:

- Facilitate science by fostering information sharing, strategy development, and communication among scientists and across disciplines to accomplish synergies, innovation, and increased integration of scientific knowledge.
- Help reduce duplication of scientific effort.
- Increase public awareness of the importance of understanding microbial systems and their capabilities.

In our work with interdisciplinary teams assembled by BER to hold discussions and develop scientific and programmatic strategies to accelerate GTL science, we create internal documentation Web sites that organize draft texts, presentations, graphics, and supplementary materials and links. From such team activities arose a number of important documents including more than 20 texts and presentations since October 2000:

- Roadmap and Web site, April 2001.
- Handouts for several BER and OASCR advisory committee meetings.
- Workshop reports.
- Numerous overview documents, including abstracts and flyers.
- Contractor-grantee workshop research abstracts book.

All GTL publications are on the public Web site. The GTL site also includes an image gallery, research abstracts, and links to program funding announcements and individual researcher Web sites. Site enhancements are under way.

In addition to the GTL Web site, we produce such related sites as Human Genome Project Information, Microbial Genome Program, Microbial Genomics Gateway, Gene Gateway, Chromosome Launchpad, and the CERN Library on Genetics. Collectively, HGMIS Web sites receive more than 10 million hits per month; one million text file hits from more than 270,000 user sessions that last an average of more than 12 minutes-well over the average time for Web visits. We are leveraging this Web activity to increase visibility for the GTL program.

GTL Communication

HGMIS also identifies venues for special GTL symposia or presentations by program managers and grantees. We present the GTL program via our exhibit at meetings of such organizations as the American Association for the Advancement of Science, American Society for Microbiology, American Chemical Society, and the Biotechnology Industry Organization, as well as the G8 energy ministers' conference hosted by DOE Secretary Abraham.

As HGMIS anticipates communications needs and new avenues to more comprehensively represent GTL science, we continually seek ideas for extending and improving communications and program integration efforts. We welcome suggestions and input.DOEGenomesToLife.org

865-576-6669

This research sponsored by Office of Biological and Environmental Research, U.S. Department of Energy. Oak Ridge National Laboratory (ORNL) is managed by UT-Battelle, LLC, for the U. S. Department of Energy under Contract No. DE-AC05-00OR22725.

Modeling/Computation

A26

Hierarchical Organization of Modularity in Metabolic Networks

Albert-László Barabási[1] (alb@nd.edu), Zoltán N. Oltvai[2] (zno008@nwu.edu), A. L. Somera[3], D. A. Mongru[3], G. Balazsi[3], Erzsebet Ravasz[1], S. Y. Gerdes[4], J. W. Campbell[4], and A. L. Osterman[4]

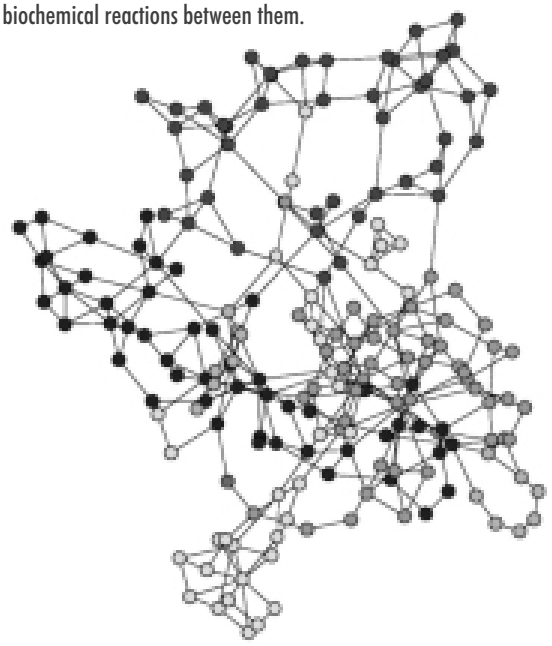
[1]University of Notre Dame, Department of Physics, 225 Nieuwland Science Hall, Notre Dame, IN 46556, 574-631-5767, Fax: 574-631-5952; [2]Department of Pathology, Northwestern University Medical School, Ward Bldg. 6-204, W127, 303 E. Chicago Ave., Chicago, IL 60611, 312-503-1175, Fax: 312-503-8240; [3]Northwestern University; and [4]Integrated Genomics, Inc.

The identification and characterization of system-level features of biological organization is a key issue of post-genomic biology. An elegant proposal addressing the cell's functional architecture is offered by the concept of modularity, assuming that the cell can be partitioned into a collection of modules. Each module, a discrete entity of several elementary components, performs an identifiable biological task, separable from the functions of other modules. Yet, it is now widely recognized that the thousands of components of the metabolism are dynamically connected to one another, such that the cell's functional properties are ultimately encoded into a complex metabolic web of molecular interactions. Within this network, however, modular organization and clear boundaries between sub-networks are not immediately apparent. Indeed, recent studies have demonstrated that metabolic networks have a scale-free topology. A distinguishing feature of such scale-free networks is the existence of a few hubs, highly connected metabolites such as pyruvate or CoA, which participate in a very large number of metabolic reactions. With a large number of links, these hubs integrate all substrates into a single, integrated web in which the existence of fully separated modules is prohibited.

To resolve the apparent contradiction, we now provided evidence that the metabolism has a hierarchical organization, an architecture that seamlessly integrates a scale-free topology with an inherent modular structure. For this purpose we have shown that the degree of clustering present in the network can be used as a distinguishing feature of a hierarchical structure, and offered direct evidence that the metabolism of 43 organisms have such a hierarchical architecture.

To turn this new conceptual framework into a practical tool we developed a method to directly identify and visualize the topological modules present in the *E. coli* metabolism and identified the function of these modules based on the predominant biochemical class of the substrates they belong to, using the standard, small molecule biochemistry based classification of metabolism. We find that most substrates of a given small molecule class are distributed within the same identified module and correspond to relatively well-delimited regions of the metabolic network, demonstrating strong correlations between shared biochemical classification of metabolites and the

Barabási— Fig. 1. The *E. coli* metabolic network color-coded based on the biochemical classification of the individual substrates. Each node corresponds to a metabolite, and links represent biochemical reactions between them.



global topological organization of *E. coli*. These results and the systematic experimental corroboration of this framework by global transposon mutagenesis will be discussed.

Supported by the DOE grant "The Organization of Complex Metabolic Networks." Principal Investigator: Albert-László Barabási, University of Notre Dame.

Background Literature

Hierarchical Organization of Modularity in Metabolic Networks, E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, *Science* Aug 30 2002: 1551-1555.

Experimental and System-Level Analysis of Essential and Dispensable Genes in *E. coli* MG1655, S.Y. Gerdes et al, in preparation.

A30

SimPheny: A Computational Infrastructure Bringing Genomes to Life

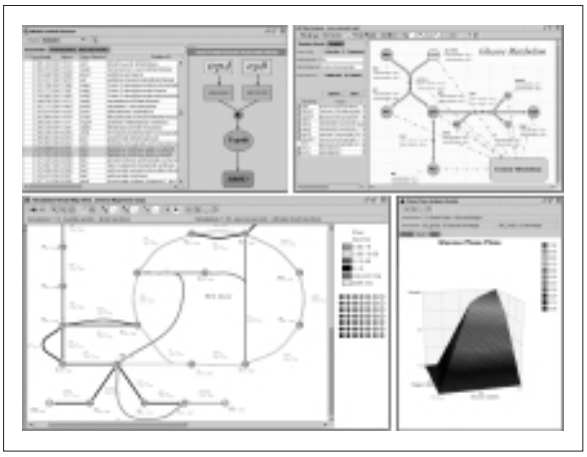
Christophe H. Schilling[1] (cschilling@genomatica.com), Radhakrishnan Mahadevan[1], Sung Park[1], Evelyn Travnik[1], Bernhard O. Palsson[2], Costas Maranas[3], Derek Lovley[4], and Daniel Bond[4]

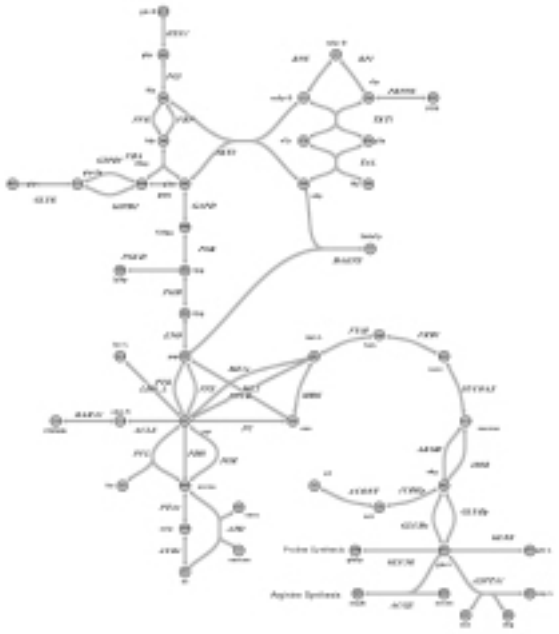
[1]Genomatica, Inc., 5405 Morehouse Drive , Suite 210, San Diego, CA 92121, 858-824-1771, Fax: 858-824-1772; [2]University of California, San Diego; [3]Penn State University; and [4]University of Massachusetts, Amherst

The Genomes to Life (GtL) program has clearly stated a number of overall goals that will only be achieved if we develop "a computational infrastructure for systems biology that enables the development of computational models for complex biological systems that can predict the behavior of these complex systems and their responses to the environment." At Genomatica we have developed the SimPheny™ (for Simulating Phenotypes) platform as the computational infrastructure to support a model-driven systems biology research paradigm. SimPheny enables the efficient development of genome-scale metabolic models of microbial organisms and their simulation using a constraints-based modeling approach.

We are currently utilizing this platform for a number of DOE-related projects including:

1. Developing the next generation of genome-scale models: In collaboration with Prof. Costas Maranas at Penn State University and Prof. Bernhard Palsson at the Univ. California, San Diego, we are integrating methods to incorporate regulation and signal transduction mechanisms into metabolic models and enable advance simulation algorithms that utilize mixed-integer linear programming (MILP).
2. *Geobacter sulfurreducens* Modeling: As part of the Microbial Cell Project led by Prof. Derek Lovley at the Univ. of Massachusetts we have completed the development of a first draft genome scale model for *G. sulfurreducens* within SimPheny. We are now beginning the process of performing simulations with the model to provide model-driven analysis of experimental data, and providing data integration solutions through the development of a model centric database
3. *Pseudomonas fluorescens* Model Development: As part of a Phase I Small Business Innovative Research (SBIR) grant we are constructing a genome scale model of *P. fluorescens* that will be used to drive metabolic engineering research on this organism for industrial bioprocessing applications.





Herein we will highlight the capabilities of the SimPheny platform as an infrastructure for supporting model driven systems biology research with special emphasis on its application to the development of the *G. sulfurreducens* metabolic model.

A32

Parallel Scaling in Amber Molecular Dynamics Simulations

Michael Crowley, Scott Brozell, and **David A. Case** (case@scripps.edu)

Dept. of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037

Large-scale biomolecular simulations form an increasingly important part of research in structural genomics, proteomics, and drug design. Popular modeling tools such as Amber and CHARMM are limited by both state-of-the-art hardware capabilities and by software algorithm limitations. Current macro-molecular systems of interest range in size to several hundred thousand atoms, and current simulations generally simulate one to tens of nanoseconds. With a 2 fs timestep, and each force evaluation involving millions of interactions to be calculated, a simulation requires many gigaflops to finish in a reasonable period of time. A parallel implementation of the calculation can provide the required performance by using the power of many processors simultaneously. However, communication speed between nodes has not progressed as rapidly as CPU processing power in recent years. Here, we address some weakness of the current parallel molecular dynamics implementation in Amber (and in a comparable program such as CHARMM). The work is aimed at making affordable a new generation of increasingly sophisticated biomolecular simulations.

Atom-Based Decomposition in Amber

Of the many ways to distribute the work of a force calculation in parallel [1,2], the method of replicated data (or "atom decomposition") has traditionally been used in Amber and CHARMM. This sort of parallel implementation is based on dividing each portion of the force calculation evenly among the processors, while keeping a full set of coordinates on all processors. This is very flexible, and relatively straightforward to program. Each processor is assigned an equal number of bonds, angles, dihedrals, and nonbond interactions. In this way, the work is balanced in each part of the force calculation, and the computation time scales well as the number of processors increases. However, in each part of the force calculation a node computes forces for different subsets of atoms. For this reason, each processor requires a complete set of up-to-date coordinates and is assumed to have components of forces for all atoms. At each step, the forces computed for all atoms on each node must be summed and distributed, and updated coordinates must be collected from each node and sent complete to all nodes. There are hence two all-to-all communications at each step. Even with binary tree algorithms for distributed sums and redistribution, the communication time becomes a significant fraction of the total time by 32 processors, even on the most sophisticated parallel machines. This limitation eliminates the possibility of efficient parallel runs at large numbers of processors, and puts a restriction on the size and length of simulations that a researcher can attempt even when large parallel computational resources are available. Still, for systems up to about 32 processors, these codes are more efficient for typical solvated simulations than are popular alternatives such as CHARMM or NAMD.

Spatial Decomposition in Amber

The second-generation parallel Amber, now under development, implements a "spatial decomposi-

tion" method [1,2] in which the molecular system is divided into regions of space where approximately equal amount of force computation is required. The method works when contributions to the force on an atom come primarily from interactions with other atoms that are relatively close and are neglected for atoms that are beyond a fixed cutoff. (This condition is valid in modern MD simulations except for long-range electrostatics, which use Ewald-based methods discussed below.) In this approach, a processor is assigned the atoms located in a slice of space and it is responsible for the coordinates, forces, velocities, and energetic contributions of those atoms. In order to compute the forces for its *owned* atoms, the processor must be able to compute the contributions from interactions with atoms that are within the cutoff, including any that are assigned to other processors. A processor keeps a copy of all such *needed* atom coordinates and forces as well as its *owned* atom coordinates and forces. At each step, a processor determines the force contributions due to all interactions in its owned and needed atoms. It sends all force contributions on *needed* atoms to the processors that own those atoms and receives any force contributions for its *owned* atoms that were calculated by other processors. When the force communications are complete, the coordinate integration is performed on the *owned* atoms. Each message in all the above communications is at most the size of the *owned* atom partition and will often be considerably smaller. Inventories of message sizes shows a reduction in overall data transferred to less than half of that for the replicated-data method, fort typical solvated protein or nucleic acid systems. Timing for communication is reduced by nearly identical ratios.

This conversion of the Amber codes is complex, since there are complications inherent in spatial decomposition that do not arise in the replicated data method; these are mainly in the treatment of bonded interactions, constraints, long-range electrostatics, and bookkeeping. The first two complications arise when molecules (chemical bonds) or distance constraints span the spatial boundaries. Most bonds, angles, dihedrals, restraints, and constraints can be assigned according to ownership of atoms. When the atoms involved are owned by distinct processors, an algorithm must be implemented to insure that the interactions are considered but only once, and that the coordinates necessary are current and correct. Bond-length constraints (using the so-called "SHAKE" approach) are more complicated, since they redefine the positions of atoms after the computed forces have been applied to owned atoms. In this case, the updated positions of all atoms involved in a constraint must be known in order to adjust positions of owned atoms regardless of whether they are owned or not. Besides these complications lie the bookkeeping needed to keep track of which forces and coordinates are being sent and received. One of the challenges is to keep the bookkeeping to a minimum, and to make it as efficient as possible, so that it does not simply replace the time saved in communications. Finally, we must optimize scaling of the Ewald method of treating long-range electrostatics in periodic system, and in particular, the PME implementation of Ewald sums. We are currently exploring several methods of reducing the communications costs of PME in highly parallel systems.

Running Dynamics on Pentium Clusters

In molecular dynamics simulations the calculations of the nonbonded interactions are a computational bottleneck. These interactions depend upon the interparticle distances. On Intel 32 bit architectures (IA32) the fastest methods to evaluate reciprocal square roots utilize the Streaming SIMD (Single-instruction multiple-data) Extensions for double precision (SSE2) operations. Tuning the AMBER source code to enable automatic vectorization, i.e., generation of SSE2 instructions by compilers, introduces a memory cache penalty as well as additional loop overhead. On IA32 platforms the SSE2 tuned AMBER is approximately eight percent faster than the original AMBER as measured by execution times of a typical explicit solvent protein simulation. The IA32 SIMD vector length is four for single precision data and two for double precision. Conversion of AMBER to single precision yields a forty percent performance improvement on IA32 platforms, as measured by execution times of a typical Generalized Born implicit solvent protein simulation. We are exploring what how to make best use of these sorts of gains without sacrificing any essential precision in the results.

[1] S. Plimpton. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117,** 1-19 (1995).

[2] L. Kalé, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, and K. Schulten. NAMD2: Greater scalability for parallel molecular dynamics. *J. Comput. Phys.* **151**, 283-312 (1999).

A34

Microbial Cell Model of *G. sulfurreducens*: Integration of in Silico Models and Functional Genomic Studies

Derek Lovley[1] (dlovley@microbio.umass.edu), Maddalena Coppi[1], Daniel Bond[1], Jessica Butler[1], Susan Childers[1], Teena Metha[1], Ching Leang[1], Barbara Methé[2], Carol Giometti[3], R. Mahadevan[4], C. H. Schilling[4], and B. Palsson[4]

[1]Department of Microbiology, University of Massachusetts, Amherst, Amherst, MA; [2]The Institute for Genomic Research, Rockville, MD; [3]Biosciences Division, Argonne National Laboratory, Argonne, IL; and [4]Genomatica, Inc., San Diego, CA 92121

Molecular analyses have demonstrated that *Geobacter* species are the predominant dissimilatory metal-reducing microorganisms in a variety of subsurface environments in which metal reduction is an important process, including uranium-contaminated aquifers undergoing bioremediation. The long-term objective of this study is to develop comprehensive conceptual and mathematical models of *Geobacter* physiology and its interactions with its physical-chemical environment, in order to predict the behavior of *Geobacter* in a diversity of subsurface environments. Prior to sequencing and initiating the functional analysis of the genomes of *Geobacter sulfurreducens* and *Geobacter metallireducens*, it was considered that these organisms were non-motile, strict anaerobes with a simple metabolism that required little regulation. It is now clear that each of these basic characterizations is incorrect. These and other surprises from the analysis of the *Geobacter* genomes are significantly influencing our design of strategies for *in situ* metals bioremediation.

A preliminary genome-scale metabolic model of *G. sulfurreducens* central metabolism has been developed with a constraints-based modeling approach using the annotated genome sequence and scientific literature. A detailed description of this model is provided in the companion poster presented by Genomatica.

Further experimental investigation is required to refine and expand this preliminary *in silico* model. For example, *G. sulfurreducens* requires a fumarate reductase in order to grow with fumarate as the sole electron acceptor, but when growing with Fe(III) as an electron acceptor, it also requires a succinate dehydrogenase in order to complete the oxidation of acetate to carbon dioxide via the TCA cycle. Although the genome of *G. sulfurreducens* contains a cluster of three genes, *frdA, frdB*, and *frdC*, resembling the three subunits of the *Wolinella succinogenes* fumarate reductase, genes for a separate succinate dehydrogenase are not apparent. Studies with a *frdA* knock-out mutant revealed that this complex functioned *in vivo*, not only as a fumarate reductase, but also as a succinate dehydrogenase. Elucidation of this novel function significantly improves the *in silico* model.

A genetic study on putative hydrogenase genes revealed that although *G. sulfurreducens* has at least four possible hydrogenases, only one is required for growth with hydrogen as the sole electron acceptor. Interestingly, this is the one hydrogenase gene that is missing from the closely related *Geobacter metallireducens*, which unlike *G. sulfurreducens* can not grow with hydrogen as an electron donor. Genetic and metabolomic studies have demonstrated that a novel cytoplasmic enzyme, which had previously been identified as NADPH-dependent Fe(III) reductase, is in fact involved in electron transport from NADPH to a variety of electron acceptors and plays a key role in intracellular NADPH homeostasis. In addition to providing valuable information on *Geobacter* metabolism, this result underscores the importance of investigating the function of enzymes *in vivo*, especially when dealing with electron transport to Fe(III), which many enzymes can nonspecifically reduce *in vitro*.

The *Geobacter* genomes contain a much higher percentage of genes devoted to electron transport function than found in other organisms. Gene expression and proteomics studies have revealed specific genes that are associated with Fe(III) reduction. For example, genes with high homology to flagellar and type(IV) pilus genes were specifically expressed during growth on the insoluble Fe(III) and Mn(IV) oxides. This finding, coupled with the discovery that *Geobacter* is chemotactic to Fe(II), has suggested that *Geobacter* produce extracellular appendages for motility only when they need to search for insoluble Fe(III) or Mn(IV) oxides (Childers, S. E., S. Ciufo, and D. R. Lovley. 2002. *Geobacter metallireducens* access Fe(III) oxide by chemotaxis. *Nature* 416:767– 769). Studies on mutants missing key pilin genes have further demonstrated the important role of pili in Fe(III) oxide reduction. In a similar manner, genes that are homologs of

components of the type II secretion systems of other bacteria are specifically expressed during growth on Fe(III) or Mn(IV) oxide and disrupting these genes inhibits the reduction of the insoluble metal oxides, but not soluble electron acceptors, including Fe(III) citrate. Mutants with deletions in one of several *c*-type cytochrome genes that are more highly expressed during growth on Fe(III) no longer had the capacity for Fe(III) reduction. These studies clearly demonstrate that *Geobacter* possess a highly regulated system of genes that are specifically involved in the reduction of Fe(III) and Mn(IV).

Additional examples, of ongoing iterative modeling and experimental elucidation of metabolic and electron transport pathways will be presented. These include novel enzymes for carbon metabolism as well as the surprising finding, originating from genome-based modeling, that *Geobacter* species that were previously classified as "strict anaerobes" have the ability to use oxygen. This has important implications for their survival under the fluctuating redox regimes common in the subsurface. Genome analysis has also suggested additional bioremediation capabilities such as the ability to degrade TNT and related contaminants and the ability to reduce mercury. These early results under the Microbial Cell Project demonstrate the power of genomic analysis to significantly influence bioremediation research.

A36

Towards a Self-Organizing and Self-Correcting Prokaryotic Taxonomy

George M. Garrity[1] (garrity@msu.edu) and Timothy G. Lilburn[2]

[1]The Bergey's Manual Trust, Michigan State University, East Lansing, MI 48224; and [2]The American Type Culture Collection, Manassas, VA 20110

A longstanding goal of microbiologists has been the creation of a taxonomy that reflects the natural history of prokaryotes and provides a stable and reliable scheme that is predictive and workable in a wide variety of applications. For genomic comparisons the establishment of such a framework is essential if we hope to be able to recognize homologous, paralogous and xenologous sequences. Over the past 15 yrs, a reasonably good picture of the evolutionary relationships among the *Bacteria* and *Archaea* has emerged, largely as a result of the widespread adoption of 16S rRNA as the molecule of choice for phylogenetic studies. At present, two large-scale phylogenetic trees of the prokaryotes exist in varying states of completeness and serve as the foundation of the comprehensive taxonomy used in the *Second Edition of Bergey's Manual of Systematic Bacteriology*. However, limitations of the underlying phylogenetic models preclude incorporation of much of the available sequence data into an all-inclusive taxonomy that can be easily maintained and modified in an automatic fashion as new taxa are described and existing taxa emended. Confounding problems include a high number of annotation errors in the sequence data and a lack of clear criteria for defining taxon boundaries.

We recently described the application of techniques drawn from the field of exploratory data analysis that take advantage of the large number of SSU rRNA sequences available and that produce results which are reconcilable with our knowledge of both the phylogenetic models and available phenotypic data. We found that principal components analysis (PCA) of large matrices of evolutionary distance data ($> 2\times10^6$ data points) yielded 2D and 3D maps of evolutionary space in which the high level groups that were formed proved consistent with those found in the large scale consensus trees. While PCA maps proved useful in establishing a comprehensive prokaryotic taxonomy and greatly aided in the identification of misclassified sequences, the method proved less useful in establishing group membership of misclassified or unknown sequences. In order to eliminate some of these problems, we have turned to a second visualization technique, heat maps. Heat maps are a type of graph in which signal intensity is displayed as a gradation of color within the confines of a precisely ordered grid. Heat maps have recently found widespread application in the field of microarray analysis and provide a useful means of visualizing differential gene expression. Heat maps, in an earlier variation, have also found application in the distant past in prokaryotic taxonomy.

Recently, we have begun using heat maps as a graphical tool for comparing alternative phylogenies and models at intermediate taxonomic levels (Order-Genus). In this paper, we describe how heat maps were used as a graphical tool to demonstrate significant improvements in the current classification of the *Gammaproteobacteria,* brought about by the application of a newly developed supervised-clustering

algorithm. Using a set of simple statistical criteria for group membership, the algorithm iteratively reorganizes the distance matrix on a taxon-by-taxon basis, excludes outliers and misidentified sequences, and subsequently reinserts such sequences into the matrix at the location of its most-probable identity. Dynamically reordered heat maps of user-selected submatrices serve as an aid to the inspection and modification of the automatically generated classification, the identification of possible classification errors, ad-hoc testing of alternative classifications/hypotheses and direct extraction of the underlying distance data.

Our results demonstrate that significant improvements to prokaryotic taxonomy can be readily obtained using statistical approaches to the evaluation of sequence-based evolutionary distances. Errors in curation, classification, and identification can be easily spotted and their effects corrected, and the classification itself can be modified so that the information content of the taxonomy is enhanced. Furthermore, evolutionary analyses based on other molecules can be viewed in terms of this rRNA-based phylogeny and used to improve the classification in taxa where the information content and resolving power of the 16S rRNA molecules proves too low (e.g., *Bacillus*). Obviously, a more robust classification has greater predictive power and serves as a more reliable evolutionary framework for genome exploration. The visualization supplies an intuitive approach, in that persons with no taxonomic training can, by looking at the heat maps, see how the classification might be improved; our algorithm formalizes and automates the means used to achieve such improvements. The chief drawback of the approach is that groups formed from taxa that are sparsely represented in the SSU rRNA sequence data set may not be as robust or stable as groups from more richly populated taxa. This is especially true in instances where such taxa are equidistant to two or more otherwise unrelated taxa. However, such problems should prove transitory as the data set grows daily and the tools we are developing will allow us to maintain and expand a comprehensive prokaryotic taxonomy.

It is worthwhile noting that the usefulness of the techniques described here is not limited to bacterial taxonomy. These methods can be used to develop and improve classifications of all types. For example, functional assignment of new sequences benefits from a reliable protein classification. Data from gene expression microarrays might also be usefully classified using these techniques.

A pseudocode description of the methods used here will be available at the poster.

A38

Computational Framework for Microbial Cell Simulations

Haluk Resat[1] (haluk.resat@pnl.gov), Heidi Sofia[1], Harold Trease[1], Joseph Oliveira[1], Samuel Kaplan[2], and Christopher Mackenzie[2]

[1]Pacific Northwest National Laboratory and [2]University of Texas Medical School at Houston

The complex nature of data characterizing cellular processes makes mathematical and computational methods essential for interpreting experimental results and in designing new experiments. Although the need to develop comprehensive approaches is widely recognized, significant improvements are still necessary to bridge experimental and computational approaches. Such advances require the development of integrated sets of computational tools to achieve the level of sophistication necessary for understanding the complex processes that occur in cells. As part of this project, we have been developing a set of prototype network analysis tools and methods, and employing them to investigate the flux and regulation of fundamental energy and material pathways in *Rhodobacter sphaeroides*.

Our tool development efforts span a wide spectrum. Current prototype components are designed in such a way that, when combined later, they will form the backbone of a comprehensive microbial cell simulation environment. In particular, we have been working on developing a framework for the following areas:

Qualitative network analysis: This analysis algorithm uses the connection diagram of the cellular networks to classify and rank them according to their linkage characteristics. We have shown that the Petri net representation can be a powerful way to extract the topologies and the universal features of cellular networks. This allows for the comparison of networks to decipher the common modules.

Gene regulatory networks: We have developed an object oriented stochastic simulation software that can be used to study the expression levels of genes. Given a regulatory network and using a user defined multistate representation for gene

expression levels, our software can be used to simulate the mean gene expression levels and their fluctuations. This simulation software was applied to derive the network parameters of a recently reported synthetic genetic network.

Stochastic kinetic simulation software: It has been well established that the number of molecules of certain species in cells can be very low. This makes the stochastic representation more appropriate to study the dynamics of cellular systems. We have shown that the kinetic simulations are not only limited to biochemical reactions but any physical event such as vesicle formation can be included in kinetic simulations. Our kinetic simulation algorithm was devised and implemented in such a way that it can be used to simulate hybrid models that combine biochemical and physical events.

Imaging of bacterial cells and image recontruction: We are in the process of obtaining images of *R. sphaeroides* using electron tomography. *R. sphaeroides* cells will be chemically processed for TEM and imaged using the high resolution electron microscope at EMSL/PNNL. A different method of electron tomography will be utilized for 3-dimensional reconstruction of a bacterial cell to visualize the spatial distribution of intracellular vesicles throughout the bacterium. This will be performed using the remove operation capabilities of the TEM imaging facility at UCSD that allows us to obtain a series of tilted digital images that can later be computationally reconstructed. Reconstructed 3-D geometry of surface features and of the internal structures will later be used in spatially resolved simulation studies.

Mesh grid based simulation framework: Explicit incorporation of geometric and structural information into cellular models is important to study the effect of the local environment on transport and kinetic properties. NWGrid and NWPhys codes that are part of the large scale computational simulation framework used at PNNL have been adapted to biological systems. Obtained images of *R. sphaeroides* will define a structure upon which we can map spatially dependent quantities and will provide the basis for building spatial computational models of this microbial cell.

Similarity analysis of genomes and prediction of superfamilies: We have developed analysis and visualization software based on clustering and data integration to enable biologists to navigate through large quantities of genome sequence data and operon information for the purpose of classifying genomic sequences and assigning protein functions rapidly and efficiently. We are applying the Similarity Box analysis to *R. sphaeroides* genome sequence data. To illustrate the usefulness of our software, we show a comparative genomics analysis of the FNR superfamily, an important group of transcriptional regulator proteins in diverse bacterial species that respond to signals such as redox, nitrogen status, and temperature. We also show a genomic comparison of FNR proteins in the two ?-Proteobacterial species, *R. sphaeroides* and *Rhodopseudomonas palustris.* These two closely related but divergent prokaryotes have a partially overlapping complement of FNR proteins.

Molecular part list and network derivation: Although most of the key genes have been identified, the network describing the energy metabolism of *R. sphaeroides* is minimally understood. To improve the energy metabolism network of *R. sphaeroides* that was initially built using biochemical data, we are using the recent genome (DOE sponsored JGI) and microarray (UT-Houston) data to build a more complete molecular parts list. We are making use of the genome data by applying our similarity analysis approach to assign functionality to improve the annotation of the genome. For similar purposes, we are analyzing the microarray data using clustering methods combined with promoter sequence information.

A40

Characterization of Genetic Regulatory Circuitry Controlling Adaptive Metabolic Pathways

Harley McAdams*, Lucy Shapiro*, and Mike Laub*

*Presenters

Stanford University

In this project, an interdisciplinary team of scientists from Stanford, Harvard, and SRI International is characterizing genetic regulatory circuitry and metabolic pathways in the bacterium *Caulobacter crescentus*. The *Caulobacter* are oligotrophic organisms adapted to low-nutrient environments such as clear streams, lakes, and the open ocean and some species are found in the deep subsurface environment. The genetic circuit controlling the *Caulobacter crescentus* cell cycle is

one of the best-characterized bacterial regulatory networks. *Caulobacter* is a particularly useful model system for study of cell-cycle regulation because cell populations can be synchronized with minimum perturbation of the normal physiology of the cell. This feature has permitted a detailed molecular analysis of the *Caulobacter* life cycle and has revealed a complex regulatory network governing the cell cycle progression and morphogenesis. The results of the current project will provide a powerful base for eventually engineering situation-specific regulatory "cassettes" into *Caulobacter* cells for targeted remediation applications.

In the first year of the project, we have developed a computational method to predict *Caulobacter* operon organization, established a baseline profile of gene expression levels for the complete genome over the cell cycle using microarrays, and developed and optimized a protocol for high throughput creation of gene deletion mutants for the entire *Caulobacter* genome. We have succeeded in obtaining and visualizing unique biofilms of *Caulobacter* wild-type and mutant cells. In their natural habitat, *Caulobacter* commonly grow in biofilms. We are particularly interested in determining whether genes are expressed in biofilms that are not active under laboratory conditions. As with most newly sequenced genomes, about forty percent of *Caulobacter*'s genes are of unknown function. Determining what these genes do is an important objective of the project.

We have published the first version of the *CauloCyc* online database (www.biocyc.org) for browsing and analysis of the *Caulobacter* genome using SRI's Pathway Tools software. This combined database and software environment has powerful and unique capabilities for modeling, visualizing, and analyzing biochemical and genetic networks. Using SRI's PathoLogic program we have computationally predicted *Caulobacter's* metabolic pathways from the genome. The PathoLogic program accepts two inputs: a fully annotated genome sequence, and the MetaCyc metabolic pathway database. The process of pathway prediction involves evaluating the evidence for the presence of each pathway in the reference DB for the organism being analyzed. A pathway consists of a sequence of enzyme-catalyzed reactions. The more enzymes we find within a pathway that are encoded by the genome, the more evidence we say there is for the presence of that pathway. The resulting metabolic pathway prediction assigned 617 *Caulobacter* enzymes to their corresponding metabolic reactions in 130 predicted metabolic pathways. Now we are investigating the "holes" within these predicted metabolic pathways, and we are using microarray studies of colonies growing in diverse media plus focused Blast studies to identify the missing enzymes within the *Caulobacter* genome.

A28

Computational Elucidation of Metabolic Pathways

Imran Shah (imran.shah@uchsc.edu)

School of Medicine, University of Colorado
http://shah-lab.uchsc.edu

Elucidating the metabolic network of a living system is an important requirement for modeling its physiological behaviour and for engineering its pathways. With the availability of whole genomes it is theoretically possible to infer the presence of putative enzymes and transporters in an organism. However, piecing this information into a complete picture is still mostly a daunting manual task for at least two reasons. First, we do not have accurate and sufficient annotation of enzymatic function from sequence. Consequently, many proteins in completely sequenced microbes remain functionally uncharacterized. Second, inferring the causal biochemical connections within a metabolic network is not straightforward. We are developing a computational infrastructure to address these challenges. In earlier work we have developed a machine learning (ML) approach to improve the assignment of enzymatic function from sequence. More recently, we have developed an artificial intelligence (AI) approach for the prediction of metabolic pathways and their interactive visualization, called PathMiner. This poster will present an overview of our system and discuss some relevant results for the radiation resistant microbe, *D. radiodurans*, and the metal-reducing bacterium, *S. oneidensis* MR-1.

A42

Data Exchange and Programmatic Resource Sharing: The Systems Biology Workbench, BioSPICE and the Systems Biology Markup Language (SBML)

Herbert M Sauro (hsauro@kgi.edu)

Keck Graduate Institute, 535 Watson Drive, Claremont, CA 91711

Standards

There is now a wide variety of modeling tools available to the budding systems biologist, but until recently there has been no agreed way to exchange models between different tools. The Systems Biology Markup Language is one of two emerging standards to allow systems biology modeling and analysis packages to exchange models. SBML is an open XML-based format developed to facilitate the exchange of models of biochemical reaction networks between software packages. Currently SBML Level 1 is supported by a growing number of tools, including, Jarnac, JDesigner, Gepasi, VCell, jigCell, Cellerator, and BioSPICE (under development). Level 2 is currently under discussion with the BioSPICE group, with a final release sometime in the second quarter of 2003.

There are in addition a small number of growing repositories of SBML models now available on the web, including, http://www.sbw-sbml.org, http://www.symbio.jst.go.jp/~funa/kegg/mge.html, http://www.sys-bio.org, http://www-aig.jpl.nasa.gov/public/mls/cellerator/nb.html, http://www.gepasi.org/

Programmatic Resource Sharing

The ERATO Systems Biology Workbench (SBW) is an open source, portable (Windows, Linux, Mac OS X) framework for allowing both legacy and new application resources to share data and algorithmic capabilities. Our target audience is the computational biology community whose interest lies in simulation and numerical analysis of biological systems. SBW allows communication between processes potentially located across a network on different hardware and operating systems. SBW currently has bindings to C, C++, Java, Delphi, Python, Perl and BioSPICE, with more planned for in the future. SBW uses a simple messaging system across sockets as a means for applications to communicate at high speed.

Software components that implement different functions (such as GUIs, model simulation methods, analysis methods, database interfaces, etc.) can be connected to each other through SBW using a straightforward application programming interface (API). The figure illustrates the visual design tool, JDesigner, interacting with the computational engine, Jarnac, via SBW (Jarnac is acting as a server and is thus invisible to the user). This setup permitted us to avoid writing, 'yet-another-simulator', and allowed JDesigner, which had no inherent simulation capabilities to dispatch simulation requests to another resource. Under SBW, models are exchanged between resource applications using SBML.

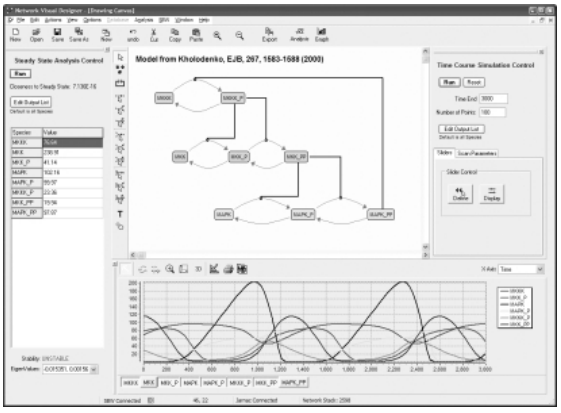
We are also working closely with the BioSPICE DARPA funded program to enhance SBML to Level 2 and to establish a set of agreed resource APIs to enable plug and play resources for both BioSPICE and SBW.

There is a growing list of modules which can communicate with each other via SBW and the BioSPICE software, including time course simulators, stochastic simulators (at least three kinds), basic optimizer, structural analysis tool via METATOOL, graphing tools, system browsing tools, etc.

Collaborators

The development of SBML/SBW was primarily conducted at Caltech by Hiroaki Kitano, John Doyle, Hamid Bolouri, Mike Hucka and Andrew Finney and Herbert Sauro in collaboration with

Sauro - Fig. 1. JDesigner simulating a MAPK Pathway



many other groups, including (in alphabetical order): A. Arkin, B. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, D. Fell, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. Juty, J. Kasberger, A. Kremling, U. Kummer, N. Le Novere, L. Loew, D. Lucio, P. Mendes, E. Mjolsness, Y. Nakayama, M. Nelson, P. Nielsen, T. Sakurada, J. Schaff, B. Shapiro, T. Shimizu, H. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, J. Wang, the BioSPICE program project investigators.

Web sites for further information on SBW and SBML:

- www.sbw-sbml.org
- www.sys-bio.org

Funding was provided by ERATO, the Keck Graduate Institute, DARPA, and a U.S. Air Force Grant.

A44

A Web-Based Laboratory Information Management System (LIMS) for Laboratory Microplate Data Generated by High-Throughput Genomic Applications

James R. Cole[1] (colej@msu.edu), **Joel A. Klappenbach**[1] (klappenb@msu.edu), Paul R. Saxman[1], Qiong Wang[1], Siddique A. Kulam[1], Alison E. Murray[2], Liyou Wu[3], Jizhong Zhou[3], and James M. Tiedje[1]

[1]Center for Microbial Ecology, Michigan State University, East Lansing, MI; [2]Earth and Ecosystem Sciences, Desert Research Institute, Reno, NV; and [3]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN

The use of high-throughput genomic-level methodologies such as microarray fabrication, proteomics, and whole-genome clone libraries necessitate computer database management tools for tracking and archiving massive quantities of data. We have developed a laboratory information management system (LIMS)—named the MicrobeArrayDB—for handing high-throughput data created during the fabrication of DNA microarrays. The laboratory microplate (96 or 384 well) functions as the primary data unit of the LIMS. This flexible database structure permits creation of different plate types, with associated data fields, extending the functionality of the LIMS to many different applications. Project-specific customization of the LIMS is controlled through a set of easily modified meta-data tables containing information on microplate types, contents, and how contents of microplates are combined and stored during laboratory procedures. The contents of microplates ("reagents") serve as "reactants" that are combined by the user during *in silico* "reactions" to create new product plates within the database. Users load microplate-specific information using cut-and-paste operations from tab-delimited file formats such as those created during oligonucleotide primer/probe design and from manufacturer supplied files. Data from these files is inherited during subsequent "reactions" that create new microplates. LIMS tools are interfaced through any internet browser and data access is controlled through group and user level permissions. User name and time stamps are recorded for each entry into the LIMS creating a permanent record and detailed audit trail. The MicrobeArrayDB is built on a multi-tier client/server architecture model using a publicly-available relational DBMS for our back-end tier (PostgreSQL) and Java web technologies for middle and presentation tiers.

In our initial project configuration, the LIMS is customized to track the production of a PCR-based DNA microarray from primer design to product deposition on coated slides. Current plate types for the microarray configuration include: Primer, Template, Primer Pair, PCR Product, and Printing plates. Contents of these plates are combined during "reactions"—such as the creation of a PCR Product plate from existing Primer and Template plates—as they are physically combined in the laboratory. Data inheritance is structured with microplate data loaded from text files produced during primer design and synthesis that include information such as gene ID/name, primer sequences, design criteria, quantity, length, batch control numbers, etc. Process information such as PCR reactions, product purification, and quality control data are added by the user during microarray construction. Internal testing has been conducted to track the fabrication process of whole-genome expression arrays for *Shewanella oneidensis* MR-1 at Oak Ridge National Laboratory. This implementation of the LIMS is targeted for whole-genome microarrays for bacteria including *Deinococcus radiodurans R1, Desulfovibrio vulgaris*, *Geobacter metallireducens*, *Nitrosomonas europaea, Rhodopseudomonas palustris.*

A46

BioSketchpad: An Interactive Tool for Modeling Biomolecular and Cellular Networks

Jonathan Webb[1], Lois Welber[1], Arch Owen[1], **Jonathan Delatizky**[1] (delatizky@bbn.com), Calin Belta[2], Mark Goulian[2], Franjo Ivancic[2], Vijay Kumar[2], Harvey Rubin[2], Jonathan Schug[2], and Oleg Sokolsky[2]

[1]BBN Technologies (http://bio.bbn.com/) and [2]University of Pennsylvania

The Bio SketchPad (BSP) is an interactive tool for modeling and designing biomolecular and cellular networks. It features a simple, easy to use, graphical front end. Descriptive models can be built and parameterized, and converted to forms supported by external simulation and analysis tools. The current version of BSP supports CHARON, a high level language and toolset for simulating hybrid systems developed at the University of Pennsylvania, and is being enhanced to communicate with other simulators, such as Virginia Tech's JigCell.

BSP was designed with biologists for biologists. It provides an intuitive graphical interface which allows experimentalists to easily generate working models of networks. Biomolecular reactions supported include transcription, translation, regulation, and general protein-protein interactions. BSP supports typical editing operations for graph editors as well as some specialized operations specific to the biomolecular modeling domain. Chemical species, reactions, and regulations of reactions are drawn as nodes in the graph. The user can control some of the rendering properties of specie elements such as the text label, color, and shape of the drawn node. Syntactic constraints are imposed on the model construction. Node highlighting is used to assist users during model construction. Specialized commands are implemented for constructing reaction geometries common to biochemical systems with specified numbers of inputs and outputs. Model parameters are specified in parameter dialogs accessed through the graphical presentation of the model. The set of parameters can change depending on which types of rate laws or regulation functions are used.

BSP is under active development. Recent and upcoming enhancements include the use of SBML Level II to exchange model information with other in silico tools, modularization of the simulator interface to utilize the standards being developed in the DARPA IPTO BioComp program, and the ability to represent models hierarchically.

BSP development has been funded by the DARPA IPTO BioComputation Program, PM Dr. Sri Kumar. The BSP application together with the CHARON simulation environment are available for download.

A48

Molecular Docking with Adaptive Mesh Solutions to the Poisson-Boltzmann Equation

Julie C. Mitchell[1] (mitchell@sdsc.edu), Lynn F. Ten Eyck[1], J. Ben Rosen[2], Michael J. Holst[3], Victoria A. Roberts[5], J. Andrew McCammon[4], Susan D. Lindsey[1], and Roummel Marcia[1]

[1]San Diego Supercomputer Center, [2]Department of Computer Science and Engineering, [3]Department of Mathematics, [4]Department of Chemistry and Biochemistry and Department of Pharmacology, University of California San Diego; and [5]Department of Molecular Biology, The Scripps Research Institute

The Docking Mesh Evaluator (DoME) uses adaptive mesh solutions to the Poisson-Boltzmann equation to quickly evaluate and optimize docking energies. This is accomplished by interpolation of potential functions over an irregular mesh that is dense in high gradient regions. The result is a method capable of performing detailed energy calculations very quickly. The initial version of DoME offers many useful tools for computational study of molecular interactions. DoME is intended to bridge the gap between methods that use a coarse interaction model for computational efficiency and those having detailed but expensive calculations. The software is fully parallel and can run on supercomputers, clusters and linked independent workstations.

The Critical Assessment of PRedicted Interactions (CAPRI) is a CASP-inspired exercise in which the goal is to predict bound protein-protein structures given their individual crystal structures. Using the Fourier transform-based molecular program DOT [1], predictions were made for seven systems in CAPRI Rounds 1 and 2. DOT's performance is an important part of DoME's development, since DoME's energy model is meant to be a higher precision, continuous version of that

used by DOT. Ours was one of four teams (out of nineteen) submitting good predictions for three of the seven systems (the most achieved by any group.) This work will appear in a special edition of Proteins [2].

As a useful starting point for the development of DoME, we are running extensive analysis on the CAPRI systems. Part of the goal in developing DoME is to achieve accurate docking solutions without requiring as extensive a search as DOT performs. We hope to determine how fineness of the sampling affects the quality of the results, in particular at what sampling level DoME returns results with a high fraction of residue-residue contacts. The effect of local optimization of the best DoME solutions generated in the global scan is being considered, as well as local optimization (using DoME) of solutions generated by DOT's more exhaustive search. It appears that optimization can disrupt correct residue-residue contacts, but it is not yet clear whether the reason for this is biological or algorithmic. We are also implementing and testing schemes for global optimization. The aim is to find a consistent "recipe" of scanning, local and global optimization that will produce useful results for most protein-protein systems. Global optimization for docking presents some difficulties, as the variables used to parameterize the system are non-homogeneous and in some cases cyclic.

[1] J.G. Mandell, V. A. Roberts, M. E. Pique, V. Kotlovyi, J. C. Mitchell, E. Nelson, I. Tsigelny and L.F. Ten Eyck (2001), "Protein docking using continuum electrostatics and geometric fit," *Protein Engineering* **14**(2): 103—115.

[2] D.H. Law, L.F. Ten Eyck, O. Katzenelson, I. Tsigelny, V.A. Roberts, M.E. Pique and J.C. Mitchell (2002), "Finding needles in haystacks: Re-ranking DOT results using shape complementarity, cluster analysis and biological information," *Prot. Struct. Fun. Gen.* In press.

A50

Functional Analysis and Discovery of Microbial Genes Transforming Metallic and Organic Pollutants: Database and Experimental Tools

Lawrence P. Wackett
(wackett@biosci.cbs.umn.edu) and Lynda B.M. Ellis (lynda@mail.ahc.umn.edu)

Center for Microbial and Plant Genomics, University of Minnesota

It is the major thesis of the current project that much of the breadth of microbial metabolism remains uncatalogued and uncharacterized. Characterizing this metabolism represents a major task of microbial functional genomics. Moreover, there is a general inability to predict metabolic pathways when all of the necessary reactions are not found in databases. The research described here seeks to better assemble existing metabolic data, discover new microbial metabolism, and predict metabolic pathways for compounds not yet in databases.

Approximately half the chemical elements are metallic and metalloid. Microbial metabolism of many of these elements, and compounds containing them, have been poorly studied relative to common intermediary metabolism, the typical focus of functional genomic analysis. Yet, recent studies suggest that many microbes have broad abilities to transform metals, metalloid elements, and compounds containing those elements. In the current project, the web-based University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD) has been expanded to include information on microbiological interactions with 52 chemical elements. For each element, there is now a webpage with annotation on the major microbial interactions with that element, links to Medline, and access to further UM-BBD information. The information above has been coordinated with several depictions of the periodic table of the elements, one a classical table with columns and rows, and secondly a depiction of the elements in a spiral. The latter serves to cluster elements better with respect to their interactions with microbiological systems. The URLs for the relevant pages are given below:

- Biochemical Periodic Tables (overview website): http://umbbd.ahc.umn.edu/periodic/

- Traditional Periodic Table: http://umbbd.ahc.umn.edu/periodic/periodic.html
- Spiral Periodic Table: http://umbbd.ahc.umn.edu/periodic/spiral.html

The pages above have added hundreds of new linkages to UM-BBD compound pages. For example, the mercury element page has 4 links, arsenic has 9 links, and chlorine has 129 links to UM-BBD compounds, respectively.

An important facet of the current project is to discover new metabolism and functionally analyze the novel microbial enzymes and genes involved. A bioinformatic analysis has shown that on the order of one hundred chemical functional groups are found in natural products, yet only fifty are currently known to undergo microbial transformation. In this context, complete genome sequence annotation will require the identification of genes and enzymes that metabolize the full diversity of elements and functional groups that microbes act on in the environment. In the current project, we are uncovering the molecular basis of microbial metabolism, some of which has been previously unknown and uninvestigated. For example, we have investigated the metabolism of bismuth compounds, boronic acids, azetidine ring compounds, and novel organonitrogen compounds.

Another goal of the project has been to develop a tool to predict microbial catabolism, using the UM-BBD as a knowledge base. The objective is to propose one or more plausible biodegradation schemes for compounds whose metabolism is not yet known. To begin, a system for substructure searching was added to the UM-BBD, which both improved the database and was a necessary component for developing a metabolism prediction software. The metabolism prediction software is based on rules that describe fundamental microbial reactions. At present, there are 88 rules in the biotransformation rule database, each specifying the atoms and their positions in a functional group and the biotransformation reaction that they undergo. The software can now predict biodegradation pathways for a significant number of aliphatic and aromatic compounds. The prototype system has been offered to UM-BBD users to use and critique. The system will be expanded with input from our Scientific Advisory Board and the broader scientific community in the coming year.

A52

Comparative Genomics Approaches to Elucidate Transcription Regulatory Networks

Lee Ann McCue[1]* (mccue@wadsworth.org), William Thompson[1], C. Steven Carmack[1], Zhaohui S. Qin[2], Jun S. Liu[2], and **Charles E. Lawrence**[1]

*Presenter

[1]The Wadsworth Center, New York State Department of Health, Albany, NY 12201; and [2]Department of Statistics, Harvard University, Cambridge, MA 02138

The ultimate goal of this research is to delineate the core transcription regulatory network of a prokaryote. Toward that end, we are developing comparative genomics approaches that are designed to identify complete sets of transcription factor (TF) binding sites and infer regulons without evidence of co-expression. Using *Escherichia coli* as our model system, we have developed a phylogenetic footprinting technique to identify TF binding sites upstream of every operon in the *E. coli* genome. This method requires the genome sequences of several closely related species, and employs an extended Gibbs sampling algorithm to analyze orthologous promoter data. Using the promoters of 166 *E. coli* operons and a database of experimentally verified TF binding sites for validation, we have evaluated our ability to predict regulatory sites with this method, and addressed the questions of which species are most useful and how many genomes are sufficient for comparison. Orthologous promoter data from just three species were sufficient for ~75% of predicted sites to overlap the experimentally verified sites by 10 bp or more. A genome-scale phylogenetic footprinting study of *E. coli* identified 741 predictions above a threshold for statistical significance ($p < 0.05$) determined using randomized data simulations. We have also developed a novel Bayesian clustering algorithm to cluster these predictions thereby identifying 181 putative regulons, most of which are orphans—that is, the cognate TF is not known. This strategy to infer regulons utilizes only genome sequence information and is complimentary to and confirmative of gene expression data generated by microarray experiments. We are now applying these technologies to the *Synechocystis* PCC6803 genome.

A54

Predicting Genes from Prokaryotic Genomes: Are "Atypical" Genes Derived from Lateral Gene Transfer?

John Besemer[1], Yuan Tian[2], John Logsdon[1], and **Mark Borodovsky**[2] (mark@amber.biology.gatech.edu)

[1]Department of Biology, Emory University, Atlanta, GA; and [2]School of Biology, Georgia Technical Institute, Atlanta, GA

Algorithmic methods for gene prediction have been developed and successfully applied to many different prokaryotic genome sequences. As the set of genes in a particular genome is not homogeneous with respect to DNA sequence composition features, the GeneMark.hmm program utilizes two Markov models representing distinct classes of protein coding genes denoted "typical" and "atypical." Atypical genes are those whose DNA features deviate significantly from those classified as typical and they represent approximately 10% of any given genome. In addition to the inherent interest of more accurately predicting genes, the atypical status of these genes may also reflect their separate evolutionary ancestry from other genes in that genome. We hypothesize that atypical genes are largely comprised of those genes that have been relatively recently acquired through lateral gene transfer (LGT). If so, what fraction of atypical genes are such *bona fide* LGTs? We have made atypical gene predictions for all fully completed prokaryotic genomes; we have been able to compare these results to other "surrogate" methods of LGT prediction. In order to validate the use of atypical genes for LGT detection, we are building a bioinformatic analysis pipeline to rigorously test each of the gene candidates within an explicit phylogenetic framework. This process starts with gene predictions and ends with a phylogenetic reconstruction of each candidate. From the set of *bona fide* LGTs that we have identified, we will be able to determine the LGT parameters to which our gene finding programs are most sensitive (*i.e.* time scale of transfers, phylogenetic distance from transfer source, *etc*.). We are developing this pipeline using four cyanobacterial genomes as our test set: *Prochlorococcus marinus* str. MIT 9313, *Prochlorococcus marinus* subsp. Pastoris str. CCMP1378, *Synechococcus sp*. WH 8102, *Synechocystis sp*. PCC 6803, the first three of which are nearly complete genomes from DOE. From this initial analysis, we are estimating the extent and pattern of LGT in each of these genomes. We will then extend our studies to include all available genomes, both complete and nearly complete.

A56

Advanced Molecular Simulations of *E. coli* Polymerase III

Michael Colvin[1] (colvin2@llnl.gov), Felice Lightstone[1], Ed Lau[1], Ceslovas Venclovas[1], Daniel Barsky[1], Michael Thelen[1], Giulia Galli[2], Eric Schwegler[2], and Francois Gygi[3]

[1]Biology and Biotechnology Research Program, [2]Physics and Advanced Technology Directorate, and [3]Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA 94552

The goal of this project is to use advanced molecular simulation methods to improve our understanding of several key mechanisms in the *E. coli* DNA polymerase III (Pol III). Pol III is the primary replicating polymerase in *E. coli* and the holoenzyme consists of at least 10 protein subunits that carry out different functions, including template-driven DNA replication (subunit α), replicative error correction (subunit e), tethering of the Pol III complex on the DNA by a "clamp" (subunit β), and the clamp loading complex (γ, δ, δ', ψ and χ subunits). In this poster we will present several key results from this study involving simulation methods ranging from homology-based structure prediction to first principles molecular dynamics simulations.

Although the three-dimensional structure of *E. coli* polymerase III β-clamp is known, and the structure of the clamp-loading complex has been solved recently, the mechanism of loading the β-clamp onto primed DNA sites remains unclear. One of the unanswered questions is what forces act to direct DNA into the β-clamp, once it becomes opened by the clamp-loader. The crystal structures of the β-clamp, clamp-loader and the monomer of the β-clamp complexed with one of the subunits (δ) of the clamp-loader makes it clear that there are no flexible domains/subdomains of any kind that could push DNA inside the ring-shaped β-clamp. On the other hand, the experimental evidence suggests that the ATP-dependent clamp-loading reaction is very efficient, arguing against a random diffusion-based mechanism.

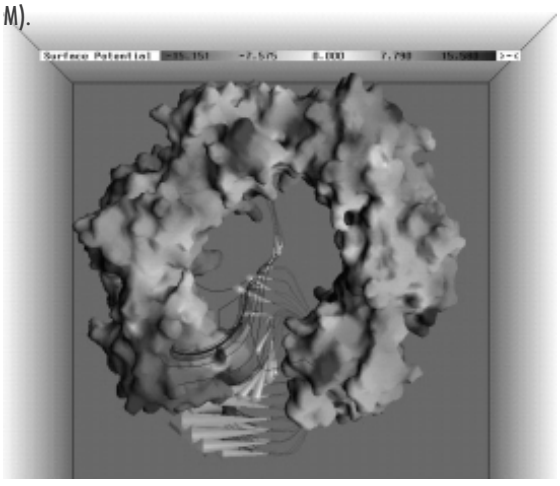
We hypothesized that the lack of the structural "helper" motifs might be compensated by other properties of the β-clamp itself, such as electrostatics. The crystal structure of a mutant β-clamp provided us with a model of the clamp opened at a single interface. Although the opening at the interface is too small for DNA to pass into the ring, we considered it a feasible model for the estimation of the effect of the electromagnetic field on the negative charges (DNA) in the vicinity of the opening. Using combination of DelPhi (a program to solve the non-linear Poisson-Boltzmann equation) and GRASP, we analyzed the electrostatic properties of the open β-clamp. The computational study revealed that the vectors of the electromagnetic field near the opened β-clamp interface are directed such that the negative electric charge (like DNA) would be drawn into the opening (see Figure). Interestingly, if the interface is opened very widely, the electrostatic guiding effect largely disappears. We are also performing classical molecular dynamics simulations of the β-subunit and its interactions with DNA.

The epsilon (ε) subunit is the 3'–5' exonuclease in this DNA polymerase and interacts with the α (polymerase unit) and θ (unknown function) subunits. Epsilon is able to hydrolyze the phosphodiester backbone of either single or double stranded DNA. This enzyme requires Mn^{2+} or Mg^{2+} for activity. The exonuclease activity resides in the N-terminus (residues 1-186). The C-terminus (residues 187-242) binds to the α subunit. The crystal structure of ε (residues 1-186) complexed with the p-nitrophenyl ester of TMP at pH 5.8 and 8.5 has recently been solved. These two structures have been provided by our external collaborator to serve as the starting point for this study. In this classical molecular dynamics study the interactions formed between a trinucleotide, modeled into the active site, and the ε subunit were investigated. Four molecular dynamics simulations were performed using explicit solvent molecules. In three separate simulations, the likely general base amino acid (His162) was modeled as a neutral residue (protonated at ND1 or NE2) and an ionized residue. The fourth simulation contained the quantum chemical gas-phase transition state docked into the active site and His162 was modeled as an ionized residue.

Our results show that the phosphodiester backbone interacts with the two Mg^{2+} ions and ε, the nucleobases form surprisingly few interactions with ε. G1 of the trinucleotide is almost completely solvent exposed. Only Met18, Asn99, and Phe102 consistently interact with the bases in all simulations. Glu61 and Ala62 interacted with the bases in the majority of the simulations. In contrast, the phosphate undergoing hydrolysis is highly stabilized in the active site. There is a minimal amount of motion in this group relative to the rest of the nucleotide. Only in the TS simulation does His162 stay close to the phosphate throughout the simulation. His162 is situated in a mobile loop which exhibits some of the highest fluctuations in the crystal structure.

In our efforts to understand the basic chemistry of the epsilon subunit (exonuclease), we have used first principles molecular dynamics simulations (FPMD) to simulate a model nuclease substrate, dimethyl phosphate (DMP) and the hydroxide-induced hydrolysis of DMP. These simulations were run on a massively parallel computer and the 14 ps DMP simulation is the longest FPMD ever reported. The results of this simulation show that the proximity of the solvated magnesium ion induces conformational changes in DMP. These subsequent conformations are higher energy conformations and could be a factor as how the enzyme cleaves the DNA backbone so easily. Hydroxide attack on DMP was simulated by fixing the distances over the reaction coordinate and then sampling for 3 ps for each constrained distance. The results show a small shoulder in the

Colvin—Fig. 1. Electrostatic surface of the slightly opened β-clamp. Lines correspond to the lines of the electric field. Arrows indicate directionality of the electric field, and their size is proportional to the calculated force. Calculations were made in solution, considering dielectric constant 2.0 for the protein interior and 80.0 for the solvent and the physiological ionic strength (0.15 M).



reaction free energy profile after the initial transition state as hydroxide attacks DMP. These results provide a reference system for subsequent simulations for the exonuclease-catalyzed reaction, and further are providing clues to answering the 40-year debate whether phosphate hydrolysis has a single transition state or has a stable intermediate.

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.

A58

Karyote®: Automated Physico-Chemical Cell Model Development Through Information Theory

Peter J. Ortoleva (ortoleva@indiana.edu), Abdalla Sayyed-Ahmad, Ali Navid, Kagan Tuncay, and Elizabeth Weitzke

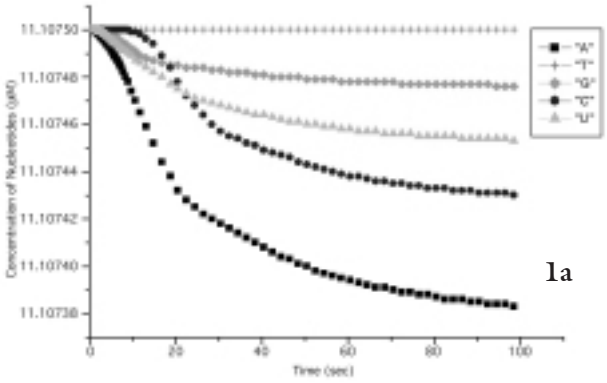
Center for Cell and Virus Theory, Indiana University

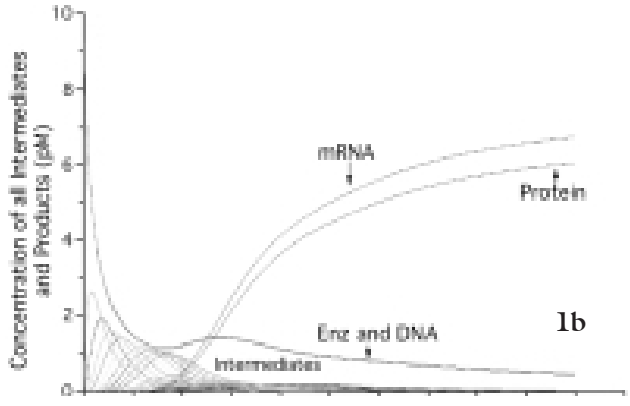
The dynamics of a cell are modeled using a reaction, transport, and genetic simulator, *Karyote*®, in order to predict cell behavior in response to changes in its surroundings or to modifications of its genetic code. Our methodology accounts for the organelles of eukaryotes and the specialized zones in prokaryotes by dividing the volume of the cell into discrete compartments. Each compartment exchanges mass with other compartments either through membrane transport or with a time delay effect associated with molecular migration (e.g. as for the nucleoid in prokaryotes). In each compartment multiple metabolic, proteomic and genomic reactions take place. All couplings among processes are accounted for. A multiple scale technique allows for the computation of processes that occur on a wide range of time scales.

Karyote® allows for the investigation of various cell behaviors that arise due to gene mutation, presence of external chemical agents or other factors. The underlying equations integrate the metabolic, proteomic and genetic networks (see Fig. 1). Catalyzed polymerization kinetics transcribes mRNA from an input DNA sequence while the resulting mRNA is used via ribosome-mediated polymerization kinetics to accomplish translation. Feedback associated with the creation of species necessary for metabolism by the genomic/proteomic network modifies the rates of production of factors (e.g. nucleotides and amino acids) that affect the genome/proteome dynamic. Hence, the effect of genetic mutations on overall cell behavior is accounted for. The concentration and sequence of the predicted proteins can be compared with experimental data via the construction of synthetic tryptic digests and associated mass spectra.

The complex network of biochemical reaction/transport processes and their biochemical spatial organization make the development of a predictive model of a living cell a grand challenge for the 21st century. However, advances in reaction/transport modeling and the exponentially growing databases of genomic, proteomic, meta-

Ortoleva—Fig. 1a and b. 1 (a) *Karyote* predicted time dependence of nucleotide concentrations during transcription for the gene TACTTTTAGGGG. As nucleotides are depleted, mRNA synthesis slows down, illustrating an important feature of coupled genome, proteome, metabolome dynamics captured in *Karyote*. (b) *Karyote* predicted evolution of concentration over time of DNA, RNA Polymerase II, all of their complexes involved in transcription/translation and mRNA created under the same condition as seen in Fig 1(a) (Weitzke and Ortoleva 2003).





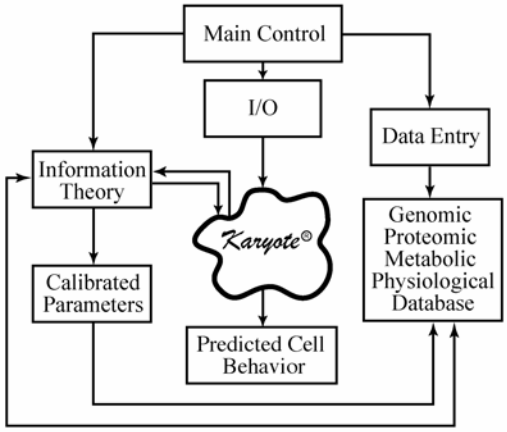


Fig. 2 This structure can serve as the basis for a national database wherein cell models are automatically developed/ calibrated and simultaneously used to interpret new data.

bolic and bioelectric data make cell modeling feasible if these two elements can be automatically integrated in an unbiased fashion. We have developed a procedure to integrate data with *Karyote®* using information theory (see Fig. 2).

Our procedure provides an objective approach for integrating a variety of types and qualities of experimental data. Data that can be used in this approach include NMR, spectroscopy, microscopy and cellular bioelectric information. The approach is demonstrated on the well-studied *Trypanosoma brucei* system.

A major obstacle for the development of a predictive cell model is that the complexity of these systems makes it unlikely that any model presently available will soon account for a complete set of processes. Thus, not only is the model-building endeavor labor intensive, but also at any stage one is faced with the challenge of calibrating and running an incomplete model. We present a probability functional method that allows the integration of quantitative and qualitative experimental data and physically motivated regularization to delineate the time course of the concentration of components (e.g. an enzyme) whose role may be key to the dynamics of the processes already incorporated in the model, but the reaction creating or destroying it are not yet understood.

A60

The Commercial Viability of EXCAVATOR™: A Software Tool For Gene Expression Data Clustering

Robin D. Zimmer[1]* (robzimmer@apocom.com), Morey Parang[2]*, Dong Xu[3], and Ying Xu[3]

*Presenters

[1]ApoCom Genomics, 11020 Solway School Road, Knoxville, TN 37931; and [2]Oak Ridge National Laboratory

ApoCom Genomics, in collaboration with Oak Ridge National Laboratory, is being funded under a DOE Phase I SBIR Grant (DE-FG02-02ER83365) to assess the commercial viability of a novel data clustering tool developed by Drs. Ying Xu, Victor Olman and Dong Xu (Xu, et.al., 2001). As we enter into an era of advanced expression studies and concomitant voluminous databases, there is a growing need to rapidly analyze and cluster data into common expression and functionality groupings. To date, the most prevalent approaches for gene and/or protein clustering have been hierarchical clustering (Eisen et.al., 1998), K–means clustering (Herwig et al., 1999), and clustering through Self-Organizing Maps (SOMs) (Tamayo et al.,1999). While these approaches have all clearly demonstrated their usefulness, they all have inherent weaknesses. First, none of these algorithms can, in general, rigorously guarantee to produce globally optimal clustering for any non-trivial objective function. Moreover K-means and SOMs heavily depend upon the 'regularity' of the geometric shape of cluster boundaries, and they generally do not work well when the clusters cannot be contained in some non-overlapping convex sets.

For cases where boundaries between clusters may not be clear, an objective function addressing more global properties of a cluster is needed. Three clustering algorithms, along with a minimum spanning tree (MST) representation, have been implemented within a computer program called EXpression data Clustering Analysis and VisualizATiOn Resource (EXCAVATOR™). Our research team has conducted a comparison between the EXCAVATOR™ clustering algorithm and the widely used K-means clustering algorithm using rat central nervous system (CNS) data. Two criteria were employed for the comparison. The first was based on the jackknife approach to assess the predictive power of the clustering algorithm,

and the second was based on the separability quality of clusters. All three of the EXCAVATOR™ algorithms (MST-hierarchical, MST-iterative, and MST-global optimal) outperformed the K-means algorithm relative to predictive power and separability quality.

In addition to comparative studies to assess the usefulness of EXCAVATOR™, the team has developed an advanced graphical user interface (GUI). The GUI has been designed to afford maximum flexibility incorporating the multi-clustering data visualization, as well as user driven comparison and editing capabilities. EXCAVATOR's™ data visualization component is based on a modular/flexible approach so as to extend its capability to other clustering/classification areas, such as phylogeny, sequence motif recognition, and protein family recognition.

References

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) *Cluster analysis and display of genome-wide expression patterns*. Proc. Natl Acad. Sci. USA, 95, 14 863-14 868.

Herwig, R., Poustka, A.J., Müller, C., Bull, C., Lehrach, H. and O'Brien, J. (1999) *Large-scale clustering of cDNA-fingerprinting data*. Genome Res., 9, 1093-1105.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999) *Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation*. Proc. Natl Acad. Sci. USA, 96, 2907-2912.

Xu, Y., Olman, V. and Xu, D. (2001) *Clustering Gene Expression Data Using A Graph-Theoretic Appraoch: An Application of Minimum Spanning*. Bioinformatics. Vol.18, no. 2002.

A62

Modeling Electron Transfer in Flavocytochrome c_3 Fumarate Reductase

Dayle M. Smith, Michel Dupuis, Erich R. Vorpagel, and **T. P. Straatsma** (tps@pnl.gov)

Computational BioSciences Group, Biological Sciences Division, Pacific Northwest National Laboratory, P.O. Box 999, MS K1-83, Richland, WA 99352, (509) 375-2802, Fax (509) 375-6631

Ferric and ferrous hemes, such as those present in electron transfer proteins, often have low-lying spin states that are very close in energy. In order to explore the relationship between spin state, geometry and cytochrome electron transfer in the flavocytochrome c_3 fumarate reductase of *Shewanella frigidimarina*, we investigate, using density functional theory, the relative energies, electronic structure, and optimized geometries for a high- and low-spin ferric and ferrous heme model complex. Our model consists of an iron-porphyrin axially ligated by two imidazoles, which model the interaction of a heme with histidine residues. Using the B3LYP hybrid functional, we found that, in the ferric model heme complex, the doublet is lower in energy than the sextet by 8.4 kcal/mol, and the singlet ferrous heme is 6.7 kcal/mol more stable than the quintet. The difference between the high-spin ferric and ferrous model heme energies yields an adiabatic electron affinity (AEA) of 5.24 eV, and the low-spin AEA is 5.17 eV. Both values are large enough to ensure electron trapping, and electronic structure analysis indicates that the iron $d\pi$ orbital is involved in the electron transfer between hemes. Mössbauer parameters calculated to verify the B3LYP electronic structure correlate very well with experimental values. Isotropic hyperfine coupling constants for the ligand nitrogen atoms were also evaluated. The optimized geometries of the ferric and ferrous hemes are consistent with structures from X-ray crystallography and reveal that the iron-imidazole distances are significantly longer in the high-spin hemes, which suggests that the protein environment, modeled here by the imidazoles, plays an important role in regulating the spin state. Iron-imidazole dissociation energies, force constants and harmonic frequencies were calculated for the ferric and ferrous low-spin and high-spin hemes. In both the ferric and ferrous cases, a single imidazole ligand is more easily dissociated from the high-spin hemes.

Environmental Genomics

B1

Identification and Isolation of Active, Non-Cultured Bacteria from Radionuclide and Metal Contaminated Environments for Genome Analysis

Cheryl R. Kuske[1] (kuske@lanl.gov), Susan M. Barns[1], and Leslie E. Sommerville[1,2]

[1]Los Alamos National Laboratory; and [2]Ft. Lewis College, Durango, Colorado

The overall goals of this project are to identify active, non-cultured soil bacteria present in radionuclide and metal contaminated environments, and to collect bacterial cells representing abundant, non-cultured populations for genome analysis. Our studies have focused on non-cultured members of the *Acidobacterium* division. Members of this division are widespread in contaminated and pristine soils having vastly different physical and chemical characteristics, and they have been found to represent a major fraction of the non-cultured bacteria in several soils (by 16S rRNA clone library analysis). Toward accomplishing the above goals we have made progress on five specific objectives.

Objective 1: Analysis of soil bacterial community rRNA. Historically, rRNA gene-based surveys of bacteria have been conducted on the pool of DNA encoding the rRNA gene. This provides a snapshot of the total composition of a sample, but does not indicate which members may be active at the time of sampling. We have conducted parallel DNA- and RNA-based analyses from a surface soil and a bacterial cell preparation extracted from that soil. By comparing the DNA-based composition to a parallel rRNA-based analysis, we hope to identify which components of the total community are active (ie. contain abundant ribosomes).

Objective 2: Survey of *Acidobacterium* division members in contaminated soils. Our second objective is to conduct a survey of DOE sites contaminated with radionuclides and metals to determine the diversity of *Acidobacterium* division bacteria and identify active members of this division in contaminated soils. Collections are in progress from the NABIR FRC (Oak Ridge, TN), PNNL, an UMPTRA site in Rifle, CO, and the Nevada Test Site. Through, we are assessing the presence and diversity of *Acidobacterium* division members using simultaneous DNA and RNA extraction, followed by 16S rRNA PCR and RT-PCR for rDNA and rRNA, respectively. Analysis of contaminated and background sites from the FRC indicate (a) subsurface biomass was very low and only DNA was recoverable in sufficient quantities for analysis from this site. (b) The composition of *Acidobacterium* division bacteria is similar across replicates of the background site, but is very different from the contaminated sites (saturated zone near wells TPB15, PTB16, DB13 in Area 2). (c) The contaminated sites contain very diverse *Acidobacterium* division members. Two new subgroups comprise the majority of 16S rRNA clones from the sample taken near well TPB16. A significant number of clones for one of the subgroups were also present in the other two contaminated sites. Clones for these new subgroups have not been found in samples from the background site.We are continuing to collect surface and subsurface samples from the other field sites for RNA- and DNA-based composition analyses.

Objective 3: Collection of *Acidobacterium* division members from soil for genome analysis. We have continued efforts to develop sensitive, specific hybridization methods for detection of *Acidobacterium* division members in environmental samples using riboprobes, and to collect hybridized cells from soil bacteria using flow cytometry cell sorting. We have developed a series of specific hybridization probes for the division and some of its major subgroups. In collaboration with Hong Cai (LANL), we determined that although cells could be specifically labeled and observed microscopically, the fluorescence intensity of each cell was too low to allow cell sorting on the LANL instrument. We are currently working with Diversa Corp. and PNNL to conduct cell sorting experiments on a Diversa instrument with better calibration for bacterial cells. At LANL, we are using gradient centrifugation to obtain

Acidobacterium division-enriched bacterial cell preparations for mixed genome libraries.

Objective 4: Culture of *Acidobacterium* division species from soils. In collaboration with Martin Keller (Diversa Corp.) and Fred Brockman (PNNL) we are using Diversa's Gel Microdroplet (GMD) technology in attempts to culture *Acidobacterium* division cells from our soil samples. Diversa has attempted culture in a matrix of different media and aeration conditions. To date we have several candidate cultures representing subgroups 1 and 6. DNA extracted from cultured cells will be used to generate genomic libraries at LANL.

Objective 5: Analysis of contaminated soil microcosm RNAs. We will use ribosomal RNA (to determine species composition) and messenger RNA (to determine active functions) to examine bacterial community response to radionuclide contamination in soil microcosm experiments. In collaboration with Mary Neu (LANL), we are setting up preliminary soil microcosms to support method development and for preliminary analysis of functional response of the natural soil bacterial community to different forms of Pu and U.

B3

A Metagenomic Library of Bacterial DNA Isolated from the Delaware River

David L. Kirchman, (kirchman@udel.edu), Matthew T. Cottrell, and Lisa Waidner

College of Marine Studies, University of Delaware, Lewes, DE 19958

Most bacteria and archaea in natural environments still cannot be isolated and cultivated as pure cultures in the laboratory, and the microbes that can be cultured appear to be quite different from uncultured ones. Consequently, the phylogenetic composition, physiological capacity and genetic properties of natural microbes have to be deduced from fluorescence *in situ* hybridization (FISH) assays and bulk properties of microbial assemblages, and from a variety of PCR-based methods applied to DNA isolated directly from natural samples. Another culture-independent approach is to clone this DNA directly into appropriate vectors and to screen the resulting "metagenomic library", which theoretically consists of all possible genes from the microbial assemblage. We applied this general approach to a sample from the freshwater end of the Delaware Estuary as part of our efforts to understand carbon and nitrogen cycling in environments like estuaries with large environmental gradients. Metagenomic libraries have been constructed for soils and marine samples, but not for freshwaters. High molecular weight DNA from the bacterial size fraction was isolated and cloned into the fosmid vector pEpiFOS-5 (Epicentre). Our library consisted of 4608 clones with an average insert size of 40 kB, representing about 90 genomes, if we assume a genome size of 2 mB.

Screening the library revealed several surprises, including genes found previously in metagenomic libraries of oceanic samples. Our library appears to be dominated by *Cytophaga*-like bacteria according to the 16S rRNA data collected by DGGE analysis of PCR amplified 16S rRNA genes. Of the 80 clones bearing 16S rRNA genes, about 50% appear to be from the *Cytophaga-Flavobacteria*, a complex cluster in the Bacteroidetes division. FISH analysis of the original microbial assemblage indicated that *Cytophaga*-like bacteria were only about 15% of the community. The next most abundant 16S rRNA genes in the library are from G+ *Actinobacteria*, which others have shown to be abundant in freshwater lakes. But beta-proteobacteria usually dominate freshwater systems and were the most abundant group in our sample according to the FISH analysis, yet beta-proteobacteria accounted for only about 15% of the 16S rRNA genes in the metagenomic library, much less than the 25% found by FISH.

The library was also screened for hydrolysis of the fluorescent analog of cellulose, MUF-beta-1,4-glucoside. Twenty-four of the 2,400 clones screened had cellulase activity, which was inferred from rates of analog hydrolysis 2.5-fold greater than the control with vector alone. The activity of seven clones was 3-fold greater than the control, while 40 additional clones had activities between 2 and 2.5-fold higher than the control. The variation of activities observed suggests that the library contains genes encoding variety of glycosyl hydrolases capable of cleaving this fluorgenic analogue. One of the cellulase-active clones was determined to harbor a 16S rRNA gene from a *Cytophaga*-like bacterium. This clone is now being completely sequenced by John Heidelburg (TIGR).

We also screened the library for genes indicative of two newly-discovered photoheterotrophic

metabolisms. Our fosmid library does not appear to contain the proteorhodopsin gene, which had been found by Beja et al. (Science (1999) 289: 1902-1906) in a marine metagenomic library. However, we found two clones that contain *pufM* and *pufL*, which code for reaction center proteins in bacteria carrying out anoxygenic photosynthesis. Although well known to be present in anoxic environments, the biophysical evidence of Kolber et al. (Science (2001) 292: 2492-2495) indicates that aerobic anoxygenic photosynthesizing bacteria are also present and perhaps are biogeochemically important in the oxic habitats, such as the oceans. One 33 kb clone from our library has a *pufL* sequence most similar to the protein sequence of the gamma-proteobacterium *Allochromatium vinosum*, whereas *pufM* of this clone is most similar to the freshwater beta-proteobacterium *Rhodoferax fermentans*. The second fosmid clone contains a 35 kb insert with the *Puf*L-M reaction center complex of alpha-proteobacterial origin. The *Puf*L sequence is most similar to that of alpha-4 proteobacteria subgroup, whereas the *Puf*M protein sequence is related to Monterey Bay environmental BAC clones and Monterey Bay isolates in the alpha-proteobacterial *Roseobacter* clade.

The diversity of rRNA genes, enzyme activities and *puf* genes in this freshwater library is consistent with our expectation that freshwater environments harbor diverse assemblages of microbes. The large number of clones with *Cytophaga*-like 16S rRNA sequences and with apparent glycosyl hydrolase activity is encouraging. Additional screening of hydrolase-active clones by fosmid-end sequencing may provide additional links to clones representing *Cytophaga*-like bacteria, which would support our efforts to explore the hydrolytic capabilities and DOM cycling by this important group of aquatic bacteria.

B5

Approaches for Obtaining Genome Sequence from Contaminated Sediments Beneath a Leaking High-Level Radioactive Waste Tank

Fred Brockman[1] (fred.brockman@pnl.gov), Margaret Romine[1], Kristin Kadner[2], Paul Richardson[2], Karsten Zengler[3], Martin Keller[3], and Cheryl Kuske[4]

[1]Pacific Northwest National Laboratory; [2]DOE Joint Genome Institute; [3]Diversa Corporation; and [4]Los Alamos National Laboratory

The SX Tank Farm at the U.S. Department of Energy's Hanford Site was built in 1953 to receive high level radioactive waste, and consists of one million gallon enclosed tanks. The waste resulted from recovery of purified plutonium and uranium from irradiated production fuels using methyl isobutyl ketone, aluminum nitrate, nitric acid, and sodium dichromate. Between 1962 and 1969, tens of thousands of gallons of radioactive liquid leaked from tank SX-108. An extreme environment formed in the vadose zone from incursion of radioactive, caustic, and toxic contaminants and heating from the self-boiling contents of the tank. Samples contain up to 50 microCuries of Cesium-137 per gram sediment, nitrate at 1% to 5% of sediment mass, pH's to 9.8, and were heated to 50 to 70 degrees C. These samples are the most radioactive sediments studied at the DOE Hanford Site in Washington state, and we hypothesized selection would result in a relatively nondiverse community. Previous work demonstrated the presence of a low number of cultured organisms from these sediments (Fredrickson et al, to be submitted).

The vast majority of microbial diversity in environmental samples has proved refractory to cultivation and therefore genome, proteome, and metabol- omic analysis. New strategies are needed to access the repertoire of genes, proteins and metabolic capabilities embodied in the community's genomic sequences. The project goal is to demonstrate an approach to obtain genetically-linked genome sequence from members of this low-biomass community. Our initial approach was to screen sediments and enrichments for the presence of communities *dominated* by a very few microbes representing *uncultured or poorly cultured* divisions (Hugenholtz et al, 1998). The purpose of the screening was to identify an appropriate

sample for constructing a community BAC library and then performing high throughput sequencing of BAC ends to assemble >1,000 Kbp of genetically linked sequence.

To evaluate the utility of this approach we first characterized the environmental DNA from 8 sediment samples and 91 enrichments by 16S amplification with conserved primers followed by sequencing of clones at the DOE Production Genomics Facility. We also performed PCR's on the samples using primers targeting 13 divisions of uncultured or poorly cultured bacteria, and cloned and sequenced successful PCR's. The results showed (1) only one sequence out of over 9,000 clones showed a BLAST hit to uncultured or poorly cultured bacteria, (2) *Archaea* 16S sequences could not be amplified, (3) biomass was about 10^5 viable cells per gram sediment, and (4) the depth of community penetration in the sediments was poor because the detection level (80,000 copies per gram sediment) was only 2 to 3 times higher than the indigenous template concentrations (determined by competitive PCR). Never the less, sequencing showed between 3 and 15 putative genera per sediment, and two of the highly contaminated sediments were dominated by alpha and/or beta proteobacteria. Proteobacteria are not primary inhabitants of Hanford Site deep vadose zone sediments, suggesting *in situ* microbial growth on components of the waste.

Because the results showed that our initial approach for studying genomes of uncultured and poorly cultured microbes at this site was not feasible, we pursued a novel culturing approach developed by Diversa Corporation. SX-108 sediments were grouped into high rad-low nitrate; high rad-high nitrate; low rad-high nitrate; and low rad-low nitrate groups. Nearby sediments recovered from similar depths were pooled to provide an uncontaminated control sample. As a positive control, a soil sample from Los Alamos Nat'l Lab. with a natural population known to be comprised of >25% Acidobacterium cells was prepared in collaboration with Cheryl Kuske. Cells were purified and concentrated from each sample using multiple nycodenz gradient centrifugations. Single cells were encapsulated in individual gel microdroplets (gmd's) and the community reconstituted by placing gel microdroplets into a column. The community was grown under low nutrient flux conditions and gmd's sorted by flow cytometry using intrinsic forward and side scatter to detect those containing microcolonies of, on average, 50 to 200 cells. Key aspects of this technology are that it enables propagation of single organisms with extremely slow growth rates, and preserves some of the community interactions and other specific requirements needed for successful cultivation.

For the positive control sample, sorted gmd's were screened with primers specific for Acidobacterium division. 16S sequences were amplified from positive gmd's and a number of the sequences group in a phylogenetic tree with subgroup 6 of the Acidobacterium division (see poster by Cheryl Kuske and Sue Barnes). This represents the first known culturing of these bacteria. Our next steps are to extract DNA by either standard techniques or employing whole genome amplification from small numbers of cells, and perform partial genome sequencing.

A total of 14,000 gmd's with putative microcolonies have been sorted from the four sets of pooled SX-108 sediments and from the uncontaminated control. In the previous study, less than 50 isolates were obtained from plating these same samples. A subset of these gmd's will be amplified, cloned, and sequenced to compare the community structure in these samples to one another, to the previously obtained isolates, and to the 16S sequences obtained by direct extraction of DNA from the sediments. Our hypothesis is that the gmd microcolonies will represent microbes from a number of poorly cultured or uncultured divisions. One or more of the most unique microbes will be used for partial genome sequencing as outlined above.

B7

Ecological and Evolutionary Analyses of a Spatially and Geochemically Confined Acid Mine Drainage Ecosystem Enabled by Community Genomics

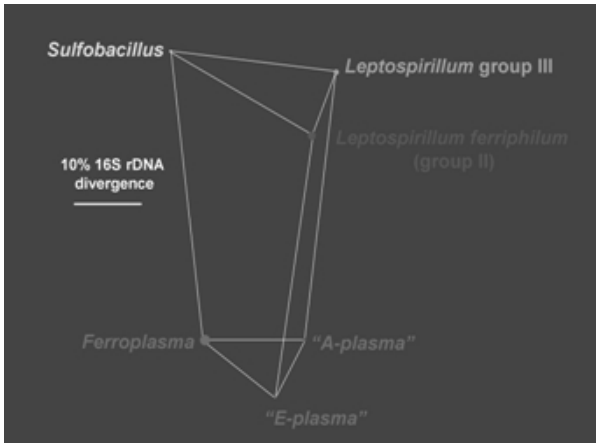
Gene W. Tyson, Philip Hugenholtz, and **Jillian F. Banfield** (jill@eps.berkeley.edu)

Department of Environmental Science, Policy and Management, Earth and Planetary Sciences, University of California, Berkeley

Subsurface acid mine drainage (AMD) ecosystems are ideal models for the genome-enabled study of microbial ecology and evolution because they are physically isolated from other ecosystems and are relatively geochemically and biologically simple.

We are using culture-independent genome sequencing of an AMD community from the Richmond mine, Iron Mountain, CA, to evaluate the extent and character of lateral gene transfer (LGT) within the community and to resolve microbial community function at the molecular level.

Microbial communities exist in several distinct habitats within the Richmond mine, including biofilms (subaqueous slime streamers and subaerial slimes) and cells attached directly to pyrite granules. All communities investigated to date by 16S rDNA clone libraries comprise only a handful of phylogenetically distinct organisms, typically dominated by the iron-oxidizing genera *Leptospirillum* and *Ferroplasma*. A *Leptospirillum*-dominated biofilm community was chosen for detailed analysis. 16S rDNA clone libraries and fluorescence *in situ* hybridization (FISH) using group-specific oligonucleotide probes indicated that the community is made up of only 6 prokaryotic populations (see Figure; the size of colored circles indicates 16S rDNA divergence within populations).



We analyzed initial community genome sequence data from a 3 Kb shotgun library of the biofilm to estimate the community genome size. This analysis used an implementation of the Lander-Waterman equation that took into consideration species abundances determined from the data and by FISH. Results indicate that each population is dominated by a single genome type. The conclusion was robust, even when assembly criteria were varied to improbable extremes and large uncertainties in population structure were included. However, more sequence data are needed to statistically validate the finding. Furthermore, the analysis is insensitive to genome types that occur in low abundance. The results suggest that reassembly of the dominant genomes of AMD community members will be tractable with a modest sequencing effort. The apparent population homogeneity may arise due to the specific characteristics of the AMD habitat or may be a widespread phenomenon in microbial ecosystems.

Similarity searches of the initial community genome sequence data revealed many genes consistent with the chemoautotrophic lifestyle of the community, including CO_2 fixation genes and nitrogen fixation genes. None of these key functional genes had close matches to genes from *Ferroplasma acidarmanus,* isolated from the Richmond Mine, the only relevant organism for which a complete genome is available. Thus, most of these genes likely come from *Leptospirillum*. Organism-resolved metabolic pathway information will be used to develop methods to monitor microbial activity in the environment.

LGT is thought to play a crucial role in the ecology and evolution of prokaryotes. The extreme conditions (pH < 1.0, molar concentrations of iron sulfate and mM concentrations of arsenic, copper and zinc, and elevated temperatures of up to 50° C) largely isolate the AMD community from most potential gene donors. Naked DNA, phage and prokaryotes native to neutral pH habitats do not persist at pH <1.0, precluding influx of genes by transformation, transduction and conjugation, respectively. However, prophage have been recognized in the *Ferroplasma* genome sequence and acidophilic phage have been detected in the biofilm community. Phage may be important vectors for gene exchange. We have initiated a collaboration to sequence the phage community to assess their diversity and enhance our ability to detect prophage in the prokaryote community genome data.

Comparative genome analyses indicate that *F. acidarmanus* and the ancestor of two acidophilic *Thermoplasma* species belonging to the Euryarchaeota have traded many genes with phylogenetically remote acidophilic *Sulfolobus* species (Crenarchaeota). The putatively transferred sets of *Sulfolobus* genes in *Ferroplasma* and the *Thermoplasma* ancestor are distinct, suggesting independent LGT events between organisms living in the same, and adjacent habitats. In both cases, however, the majority of transferred genes are involved in metabolism, particularly energy production/conversion and amino acid transport/metabolism. The lack of genes transferred from the (sequenced) genomes of other

prokaryotes is consistent with the hypothesis that extreme acidophiles have limited access to genes from organisms outside their ecotype. Interestingly, *Sulfolobus*, *Ferroplasma* and *Thermoplasma* are all bounded by a single tetraether-dominated membrane, which may facilitate conjugation. To date, no *Sulfolobus* species have been detected at Iron Mountain, suggesting two possibilities to explain the observed pattern of putatively transferred genes to *Ferroplasma* from *Sulfolobus*: 1) *Sulfolobus* is present at Iron Mountain but in regions currently inaccessible to sampling and/or 2) the transfers occurred prior to introduction of *Ferroplasma* into the current geological setting. Comparative analyses of the community genome data should improve the resolution of LGT in the community.

Ultimately, our goal is to develop an understanding of how acidophilic organisms evolved and function as communities to control acid mine drainage generation. The community genomics data are essential for this effort.

Microbial Genomics

B11

Strategies to Harness the Metabolic Diversity of *Rhodopseudomonas palustris*

Caroline S. Harwood[1] (caroline-harwood@uiowa.edu), Jizhong Zhou[2], F. Robert Tabita[3], Frank Larimer[2], Liyou Wu[2], Yasuhiro Oda[1], Federico Rey[1], and Sudip Samanta[1]

[1]The University of Iowa; [2]Oak Ridge National Laboratory; and [3]Ohio State University

Rhodopseudomonas palustris is an extremely successful bacterium that can be found in virtually any temperate soil or water sample on earth. It can grow by adopting one of the four major metabolic modes: photoautotrophic growth (energy from light and carbon from C1 compounds), photoheterotrophic growth (energy from light and carbon from organic compounds), chemohetero- trophic growth (carbon and energy from organic compounds) and chemoautotrophic growth (energy from inorganic compounds and carbon from C1 compounds). *Rhodopseudomonas* enjoys exceptional versatility within each of these growth modes. It can grow with or without oxygen and can use many alternative forms of carbon, nitrogen and inorganic electron donors. It degrades plant biomass and chlorinated pollutants and shows promise as a catalyst for biofuel production. *Rhodopseudomonas* has thus become a model to probe how the web of metabolic reactions that operates in the confines of a single cell adjusts in response to subtle changes in environmental conditions. Genes for key metabolic enzymes and regulatory proteins are easily identified in the 5.49 Mb *Rhodopseudomonas* genome. It has a large cluster of photosynthesis genes and a collection of additional genes that encode light-responsive proteins. It has genes for the metabolism of diverse kinds of carbon sources, including lignin monomers, complex fatty acids and dicarboxylic acids. It encodes two different carbon dioxide fixation enzymes and three different nitrogen fixation enzymes, each with a different transition metal at its active site. Each of the three nitrogenases is active in *Rhodopseudomonas* and each catalyzes the conversion of nitrogen gas to ammonia and hydrogen, a biofuel.

Rhodopseudomonas can convert sunlight to ATP and derive electrons by biodegrading plant material. ATP and electrons so generated can, in turn, be used to fix nitrogen, with accompanying hydrogen production. Because multiple systems are involved, hydrogen production is a good starting point for studies or integrative metabolism by *Rhodopseudomonas.* To achieve efficient hydrogen production it is important to understand how expression levels of genes involved in photosynthesis, carbon dioxide fixation, lignin monomer degradation and nitrogen fixation fluctuate in response to variations in conditions. It is also important to identify regulatory bottlenecks that restrict the flow of energy and electrons from plant biomass to hydrogen production. Towards this end we have constructed a whole genome DNA microarray of *Rhodopseudomonas* and we are now using the array to analyze global patterns of gene expression.

B13

Gene Expression Profiles in *Nitrosomonas europaea*, an Obligate Chemolithoautotroph

Dan Arp[1] (arpd@bcc.orst.edu), Xueming Wei[1], Luis Sayavedra-Soto[1], Martin G. Klotz[2], Jizhong Zhou[3], and Tingfen Yan[3]

[1]Oregon State University; [2]University of Louisville; and [3]Oak Ridge National Laboratory

Nitrosomonas europaea derives energy for growth from the oxidation of ammonia to nitrite. This process contributes to nitrification in soils and waters, often with detrimental effects in croplands, and with beneficial effects in wastewater treatment. Our long-range goal is to understand the molecular underpinnings for the oxidation of ammonia and other cellular processes carried out by these organisms. Towards this goal, the genome of this bacterium was sequenced at the Lawrence Livermore National Laboratory (Jane Lamerdin, Patrick S. Chain). Through a collaborative effort with at Oak Ridge National Lab-

oratory, we are developing microarrays in order to examine whole genome expression profiles for this organism. We are interested in understanding the effects of nutrient shifts, starvation, and other environmental changes on gene expression. The arrays are now constructed and preliminary results will be presented.

Prior to the initiation of this project, the expression of the genes coding for the enzymes involved in ammonia oxidation were characterized. These genes are *amoCAB*, coding for ammonia monooxygenase, and *hao*, coding for hydroxylamine oxidoreductase. The *amo* genes in particular show a strong response to ammonia. In preparation for the microarray experiments, we were interested in learning more about the expression of the genes coding for Rubisco. This autotroph assimilates CO_2 via this enzyme and the Calvin Cycle. Sequence data reveals that *N. europaea* has a type I Rubisco. The Rubisco operon in *N. europaea* likely consists of five open reading frames. Although Rubisco is essential for the autotrophy of this organism, studies on its expression are scant. We analyzed mRNA levels of Rubisco and some other genes by Northern hybridization with specific probes. Rubisco large and small subunits (encoded by *cbbL* and *cbbS*) were highly expressed in growing cells. The message levels appeared higher than those of *amo* and *hao*. In ammonia-deprived and stationary phase cells, Rubisco mRNAs were undetectable. Particularly interesting is that the expression of Rubisco genes was inversely proportional to the carbon levels in the medium. Rubisco mRNAs in cells grown in a medium with only atmospheric CO_2 were several times higher than those in cells in carbonate-containing medium. The higher the carbonate level in the medium, the lower the Rubisco mRNA levels. This result was in contrast to house keeping genes such as *hao* and carbonic anhydrase genes.

We also investigated the induction of the *cbb* operon and its message depletion patterns. After 30 min induction in normal medium, abundant *cbbL* and *cbbS* messages were detected and they were expressed fully after one hr. The estimated halflives of *cbbL* and *cbbS* were 0.5 and 0.75 hr, which were shorter than the half lives of *amo* and *hao*.

The three remaining *cbb* genes in the operon were expressed at much lower levels than *cbbL* and *cbbS*. This result may be explained by the presence of a transcription terminator downstream of *cbbS*, as revealed in the DNA sequence data. These results indicated that Rubisco gene expression was dependent on ammonia, and that carbon had a negative control on Rubisco transcription.

B15

Genomics of *Thermobifida fusca* Plant Cell Wall Degradating Proteins

David B. Wilson (dbw3@cornell.edu), Yuan-Man Hsu, and Diana Irwin

Department of Molecular Biology & Genetics, Cornell University, Ithaca, NY 14853

Plasmids have been successfully introduced into *Thermobifida fusca* YX by mating to *E. coli* cells containing a transfer plasmid and selecting for thiostrepton resistance. We have constructed a suicide plasmid with a defective *celR* gene containing an insert and are trying to knockout the *celR* gene in *T. fusca*. One puzzling result is that mating does not work in *T. fusca* strain ER_1, which was isolated from *T. fusca* YX by mutagenizing spores with methyl sulfate and selecting for a colony lacking the major extracellular protease. The two strains should only differ by a few point mutations and it seems unlikely that inactivating the protease would interfere with mating or plasmid transfer.

We have continued our study of *T. fusca* XG74, the main *T. fusca* xyloglucanase. Surprisingly, *T. fusca* does not grow on xyloglucan and this appears to be due to the lack of a transport system for taking up the products of xyloglucan hydrolysis since they accumulate in the media of *T. fusca* cells incubated with xyloglucan. We have shown that XG74 does allow a mixture of *T. fusca* cellulases to hydrolyze cellulose coated with xyloglucan, which is not hydrolyzed by the mixture containing only cellulases. The mixture containing XG74 cannot degrade the cellulose in tomato cell walls, but *T. fusca* crude cellulase will it. We will try to identify the additional proteins that are required for hydrolysis. The level of Xg74 in the culture supernatant of *T. fusca* grown on different carbon sources was determined by Western blotting and it was low on glucose or cellobiose slightly higher on xylan and high on corn fiber, Sulka Floc or xyloglucan.

B17

The *Rhodopseudomonas palustris* Microbial Cell Project

F. Robert Tabita[1] (tabita.1@osu.edu), Janet L. Gibson[1], Caroline S. Harwood[2], Frank Larimer[3], Thomas Beatty[4], James C. Liao[5], Jizhong (Joe) Zhou[3], and Richard Smith[6]

[1]Ohio State University; [2]University of Iowa; [3]Oak Ridge National Laboratory; [4]University of British Columbia; [5]University of California at Los Angeles; and [6]Pacific Northwest National Laboratory

The long-range objective of this interdisciplinary study is to examine how processes of global carbon sequestration (CO_2 fixation), nitrogen fixation, sulfur oxidation, energy generation from light, biofuel (hydrogen) production, plus organic carbon catabolism and metal reduction operate in a single microbial cell. The recently sequenced *Rhodopseudomonas palustris* genome serves as the raw material for these studies since the metabolic versatility of this organism makes such studies both amenable and highly feasible. Bioinformatics analysis has allowed for a reasonable approximation of many of the metabolic schemes that are utilized by this organism to catalyze the above processes. However, it appears from recent studies that several of the above processes are coordinately controlled and interdependent; thus during the first part of this project we have focused at identifying key regulatory genes and proteins. From the genomic sequence a number of likely target genes of opportunity were also revealed. These have been systematically knocked out to produce a battery of useful mutant strains that are employed in a variety of studies to examine the regulation of metabolism. In addition, we have taken advantage of a transposon library in which virtually all the open reading frames of the genome have been interrupted. Screening this transposon library under diverse growth conditions has enabled us to identify several additional unique and previously unappreciated genetic loci that are important for the above processes. The latter mutant strains are currently being studied to reveal exactly how these newly identified genes and their products influence the processes under study.

In addition to these more traditional approaches, we have also undertaken genomics, proteomics, and metabolomics oriented experiments to further analyze the integrative control of metabolism, with the rationale that these approaches will help direct our future investigations and focus our efforts. For example, whole genome microarrays have been prepared for *R. palustris* and initial studies have commenced that are assisting our efforts to identify additional genes that are up and down regulated under growth conditions of interest, with examination of both wild-type and mutant strains underway or contemplated in the near future. Furhermore, a Bioinformatics method was developed to deduce operon structure using microarray data and gene distance. This method allows the refinement of operon prediction based on genomic information. In contrast to the theoretical prediction based only on genomic information, this method incorporates microarray data and considers the noise level in the microarray experiments. In parallel with microarray experiments, examination of the proteome has commenced under selected growth conditions, using both whole cells and isolated intracytoplasmic membranes (these are intracellular structures that are employed for photochemical energy generation by this and related organisms). Proteomic analysis provides a real advantage as one can examine the end result of transcription and translation under different physiological growth conditions and use this protein data to relate back to the regulation of gene expression and/or potential posttranslational events.

The above experimental approaches will allow us to reach the eventual goal of this project; i.e., to generate the knowledge base to model metabolism for the subsequent construction of strains in which carbon sequestration and hydrogen production are maximized.

B19

Lateral Gene Transfer and the History of Bacterial Genomes

Scott R. Santos and **Howard Ochman**

Department of Biochemisty and Molecular Biophysics, University of Arizona, Tucson, AZ 85721

Comparative analyses of complete microbial sequences have brought many new insights into the evolution of genomes and the genetic relationships among microorganisms. For the vast majority of microbial species, molecular phylogenetic relationships have been on a single gene, *i.e.*, small subunit ribosomal DNA (16S rDNA). Because 16S rDNA is highly conserved – both in

terms of its function and distribution in all lifeforms and its rate of change – it is particularly well-suited for resolving the relationships among very divergent organisms. However, recent studies provide clear evidence that lateral gene transfer is common among bacteria with the result that genomes are chimeric, such that different regions will have very different histories.

The long-range objectives of our reseach is to employ nucleotide sequences of a large set of universally distributed genes among eubacteria of differing degrees of genetic relatedness in order to address questions relating to the role of gene transfer in shaping bacterial genomes. We are using existing databases as well as newly determined nucleotide sequences in order to design conserved primers for PCR amplification and sequencing of nucleotide sequences across taxa. These primer sets were initially derived from alignments of 143 protein coding genes common to the majority of eubacteria. Of these, conserved primer sets, i.e., those of limited degeneracy that yield products 400-2000 bp in length, could be obtained for 21 of these genes. Primer sets were then tested against DNA templates from representatives of diverse bacteria phyla, and the initial screening identified nine genes that could be amplified reliably from over 50% of the isolates.

The target genes for which we have developed conserved primer sets are involved in a variety of cellular functions, including translation, ribosomal structure and biogenesis (fusA, ileS, leuS, rplB, valS), DNA replication, recombination and repair (gyrB), cell motility and secretion (lepA), nucleotide transport and metabolism (pyrG) and transcription (rpoB). This diversity makes them ideal candidates to test if genes subject to lateral transfer are functionally or genetically linked and, to test the phylogenetic limits to lateral gene transfer. We are also developing primer sets for proteins specific to members of the specific phyla (e.g., *Proteobacteria*, *Spirochaetes*) to explore the ancestry of taxa-specific genes.

B21

Environmental Sensing, Metabolic Response, and Regulatory Networks in the Respiratory Versatile Bacterium *Shewanella oneidensis* MR-1

James K. Fredrickson[1] (jim.fredrickson@pnl.gov), Margie F. Romine[2], William Cannon[2], Yuri A. Gorby[2], Mary S. Weir-Lipton[2], H. Peter Lu[2], Richard D. Smith[2], Harold E. Trease[2], and Shimon Weiss[2]

[1]Pacific Northwest National Laboratory; and [2]University of California at Los Angeles

Shewanella oneidensis MR-1 is a motile facultative bacterium with remarkable metabolic versatility in regards to electron acceptor utilization; it can utilize O_2, nitrate, fumarate, Mn, Fe, and S^0 as terminal electron acceptors during respiration. This versatility allows MR-1 to efficiently compete for resources in environments where electron acceptor type and concentration fluctuate in space and time. The ability to effectively reduce polyvalent metals and radionuclides, including solid phase Fe and Mn oxides, has generated considerable interest in the potential role of this organism in biogeochem- ical cycling and in the bioremediation of contaminant metals and radionuclides. In spite of considerable effort, the details of MR-1's electron transport system and the mechanisms by which it reduces metals and radionuclides remain unclear. Even less is known regarding the molecular networks in this organism that allow it to respond to compete efficiently in a changing environment. The entire genome sequence of MR-1 has been determined and high throughput methods for measuring gene expression are being developed and applied.

DOE recognized that in order to achieve the goals of the Genomes to Life Program, obtaining a comprehensive systems-level level understanding of the components and functions of the cell that give it life, a united effort integrating the capabilities and talents of many would be required. To this end, the *Shewanella* Federation was formed to probe in detail the functions of *Shewanella oneidensis* MR1 cells. The Federation consists of teams of scientists from academia, national laboratories, and private industry (shewanella.org) that are working together in a collaborative, coordinated mode to jointly achieve a comprehensive understanding of the biology of this remarkably versatile organism. This project is contributing to

the collaborative experiments of the Federation by providing, among other things, characterized chemostat cultures of MR-1 and high resolution separation and high mass accuracy and sensitivity Fourier transform ion cyclotron resonance (FTICR) mass spectrometry for global proteome analyses of these cultures. To date, >50% of the predicted 5000+ proteins in MR-1 have been identified by accurate mass tags (AMTs) with an average of 3 AMTs per protein.

To date, we have established procedures for growing MR-1 in continuous culture under both aerobic and anaerobic (with fumarate) conditions with lactate as the growth-limiting nutrient for the initial collaborative *Shewanella* experiments. The initial results from 2-D PAGE analysis of proteins by C. Giometti (ANL) revealed that there were substantial variations in the proteome in the aerobic vs. anaerobic cultures and that biological replicates of chemostat samples were in excellent agreement. Microarray based analyses of mRNA expression are underway in collaboration with J. Zhou (ORNL) as is MS-based proteome analyses at PNNL. An unanticipated result was the formation of flocs in the aerobic chemostat. Floc formation was due to the production of exopolymeric substance (EPS) by MR-1 and was hypothesized to be due to a defense mechanism against O_2 stress (i.e., induced by O radicals). In the absence of Ca, flocs are unable to form likely due to a lack of cross- linking of the EPS. We are currently generating additional MR-1 continuous cultures in the absence of Ca under aerobic and anaerobic conditions as well as under O_2 limiting conditions. The major goal of these experiments is to provide insights into gene and protein expression patterns under these conditions as a baseline for future experiments with other types of electron acceptors, such as metals and radionuclides or nitrate, and for identifying genes that are potential targets for mutagenesis.

Another approach for characterizing differential gene expression can be provided by analysis of reporter activity mediated by transcriptional fusions. Because reporter activity can be measured in living cells in real-time, the use of transcriptional fusions is more amenable than are microarrays to dynamic measurements of gene expression analysis under many growth conditions and measurements can be made at the level of individual cells as opposed to a bulk average of the entire population. We describe the construction of a small targeted reporter library (62 constructs) in MR-1, whereby promoter-containing DNA sequences upstream to genes associated with electron transport, adhesion, and cell signaling were cloned in a broad-host range plasmid upstream to the green fluorescent protein (GFP). Using MR-1 bearing GFP reporter constructs, we have demonstrated that 1) the vector, pProbe-NT, utilized is stable after several passages in the absence of antibiotic, 2) GFP is stable for at least a week, and 3) very little time is needed for fluorescence development, even where cells are grown anaerobically prior to making fluorescence measurements. Initial testing using a crude assay developed to measure the effect of growth in suspension versus on solid surfaces, suggest that expression of the promoter upstream to *mtrDEF* is significantly higher on surfaces than in suspension. Similar, but not so dramatic, effects were observed for other promoter constructs. Proposed methods for high throughput analyses with these constructs and the utility of these assays for design of complementary microarray analyses are underway.

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. 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.

A64

Interdisciplinary Study of *Shewanella oneidensis* MR-1's Metabolism and Metal Reduction

Eugene Kolker (ekolker@biatech.org)

BIATECH (www.biatech.org), 19310 N. Creek Parkway, Suite 115, Bothell, WA 98011, 425.481.7200 x100, Fax: 425.481.5384

Since our project became part of the Shewanella Federation, we focused our work mostly on analysis of different types of data produced by global high-throughput technologies to characterize gene and protein expression as well as getting a better understanding of the cellular metabolism. Specifically, first year activities include development of:

1. New labeling technique for quantitative proteomics, so called methyl esterification labeling approach, complementary to currently available methods;
2. New algorithm for de novo protein sequencing;
3. New statistical model for spectral analysis of arbitrary shape data;
4. One of the first analyses of the transcriptome of the entire microorganism;
5. New approach to predict operon structures and transcripts within untranslated regions;
6. The first control protein experimental mixtures with known physico-chemical characteristics for high-throughput proteomics experiments;
7. The first statistical models for peptide and protein identifications for high-throughput proteomics analysis;
8. *Shewanella* metabolic capability experiments with minimal media on aerobically & anaerobically grown cells and transformation experiments.

Several collaborations have been established within the *Shewanella* Federation with PNNL, USC, ORNL, and MSU. The first year of this project, supported by DOE's Offices of Biological and Environmental Research and Advanced Scientific Computing Research, also resulted in 6 published papers.

This is a joint work of A. Keller, A. Nesvizhskii, A. Picone, B. Tjaden, D. Goodlett, S. Purvine, S. Stolyar, and T. Cherny done at BIATECH and ISB.

B23

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

Jizhong Zhou[1] (zhouj@ornl.gov), Dorothea K. Thompson[1], Matthew W. Fields[1], Adam Leaphart[1], Dawn Stanek[1], Timothy Palzkill[2], Frank Larimer[1], James M. Tiedje[3], Kenneth H. Nealson[4], Alex S. Beliaev[5], Richard Smith[5], Bernhard O. Palsson[6], Carol Giometti[7], Dong Xu[1], Ying Xu[1], Mary Lipton[5], James R. Cole[3], and Joel Klappenbach[3]

[1]Oak Ridge National Laboratory; [2]Baylor College of Medicine; [3]Michigan State University; [4]University of Southern California; [5]Pacific Northwest National Laboratory; [6]University of California at San Diego; and [7]Argonne National Laboratory

Large-scale sequencing of entire genomes represents a new age in biology, but the greatest challenge is to define cellular responses, gene functions, and 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. To define the repertoire of MR-1 genes responding to different terminal electron acceptors, transcriptome profiles were examined in batch cultures grown with fumarate, nitrate, thiosulfate, DMSO, TMAO, ferric citrate, ferric oxide, manganese dioxide, colloidal manganese, and cobalt using DNA microarrays covering ~99% of the total predicted protein-encoding open reading frames in *S. oneidensis*. Total RNA was isolated from cells exposed to different electron acceptors for 3.5 h under anoxic conditions and compared to RNA extracted from cells under fumarate-reducing conditions. Microarray analy-

ses revealed significant differences in global expression patterns in response to different anaerobic respiratory conditions. The data indicated a number of genes that displayed preferential induction in response to specific terminal electron acceptors. This work represents an important step towards the goal of characterizing the anaerobic respiratory system of *S. oneidensis* MR-1 on a genomic scale.

To understand the molecular basis of anaerobic energy metabolism in MR-1, more than 20 genes with putative functions in global gene regulation, energy metabolism and adaptive cellular responses to stress were inactivated by deletion mutagenesis. Three double mutants defective in two global regulatory genes were also obtained. Genetic, biochemical and physiological characterization of these mutants are currently underway. Also, a random *Shewanella* phage display library is being constructed, and this library will contain 10 million unique inserts with insert sizes ranging from 300 bp to about 1.2 kb. Thus, the library should have a fusion point for approximately every base pair in the genome. In addition, the conditions for cloning individual open reading frames from *Shewanella* were optimized, and now the design and synthesis of primers for amplifying individual genes are underway.

B25

Global Regulation in the Methanogenic Archaeon *Methanococcus maripaludis*

John Leigh[1] (leighj@u.washington.edu), Murray Hackett[1], Roger Bumgarner[1], Ram Samudrala[1], **William Whitman**[2], Jon Amster[2], and Dieter Söll[3]

[1]University of Washington; [2]University of Georgia; and [3]Yale University

Methanococcus maripaludis is a model hydrogenotrophic methanogenic archaeon. Growth on hydrogen and carbon dioxide results in the production of methane as a waste product. *M. maripaludis* stands out among methanogenic archaea as an ideal model species because of fast reproducible growth, a genome sequence, and effective genetic tools. Genetic manipulations in *M. maripaludis* are increasingly facile. In our collaboration under the Department of Energy's Microbial Cell Program we are studying global regulation by hydrogen, amino acid starvation, growth rate, and other conditions. Little is known of these regulatory systems in Archaea, especially at the global level. Our approaches include continuous culture of *M. maripaludis*, expression arrays, proteomics, measurement of metabolite levels, determination of tRNA charging, and genetic manipulation.

To date our team has installed dual chemostats running with anaerobic gas sources and has established reproducible growth conditions over a range of growth rates. We are in the process of calibrating the continuous culture system for growth under various nutrient limitations including hydrogen. Besides having well controlled steady-state conditions, a particular virtue of the continuous culture approach is that it will allow us to distinguish the specific effects of nutrient limitation from the general effects of growth rate. As work preliminary to our global analysis of hydrogen regulation we have produced *lacZ* fusions to *mtd* (encoding the fourth step in methanogenesis) and two *fdh* genes (encoding formate dehydrogenases) and have demonstrated differential regulation under high- and low hydrogen regimes. In preparation for our study of the response to amino acid starvation we have constructed two amino acid-auxotrophic mutants, including a mutant in the biosynthesis of leucine (isopropyl malate synthase, *leuA*) and a mutant in the common pathway of aromatic amino acids biosynthesis (3-dehydroquinate dehydratase, *aroD*). The genotypes of these mutants have been confirmed, and the mutants possess the expected auxotrophic phenotypes. We are also preparing to study other regulatory systems, and for this purpose we have constructed null mutations in genes encoding potential transcriptional regulatory proteins, including members of the AsnC, MerR and GntR families. For the analytical aspects of our global regulation studies we are implementing an approach to proteomics that in preliminary tests appears well suited to quantitative global analysis of the proteomes of prokaryotic organisms with relatively small genomes. In this approach, pools of proteolytic peptides are fractionated by several stages of liquid chromatography, analyzed by tandem mass spectrometry, and computationally matched to open reading frames in the genome. For expression studies at the mRNA level we have purchased a DNA array for *M. maripaludis* and are in the process of spotting glass slides. The expression array center at the University of Washington is continuing to develop data analysis tools that will facilitate our manipulation and integration of expression array data, proteomic data, and annotation information.

B27

Identification of Regions of Lateral Gene Transfer Across the Thermotogales

Karen E. Nelson (kenelson@tigr.org), Emmanuel Mongodin, Ioana Hance, and Steven R. Gill

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland, Telephone: 301-838-3565, Fax: 301-838-0208

The genome of *Thermotoga maritima* MSB8 was completely sequenced in 1999. Whole genome analysis of this bacterium suggested that 24% of the DNA sequence was most similar to that of archaeal species, primarily to *Pyrococcus sp.* Many of these open reading frames (ORFs) that were archaeal-like were clustered together in large contiguous pieces that stretched from 4 to 21 kb in size, were of atypical composition when compared to the rest of the genome, and shared gene order with the archaeal species that they were most similar to. The analysis of the genome suggested that this organism had undergone extensive lateral gene transfer (LGT) with archaeal species. Independent biochemical analyses by Doolittle and workers using degenerative PCR and subtractive hybridization techniques have also revealed gene transfer and extensive genomic diversity across different strains of *Thermotoga*. Genes involved in sugar transport, polysaccharide degradation as well as subunits ATPases were found to be variable.

We have created a whole-genome microarray based on the completed genome sequence that has been used to do comparative genome hybridizations (CGH) with 12 different *Thermotoga* strains/species that include *Thermotoga* sp. RQ2, *Thermotoga neopolitana* NS-ET and *Thermotoga thermarum* LA3. PCR products representing the 1879 total *T. maritima* MSB8 ORFs have been spotted in duplicate on Corning UltraGap slides. Two flip-dye experiments have been conducted per strain. Genes were considered to be shared between the 2 compared strains if the ratio (MSB8/experimental strain) was between 1 and 3, and considered to be absent if the ratio was greater than 10. Analysis of the resulting data demonstrates that there is a high level of variability in the presence and absence of genes across the different *Thermotoga* strains/species.

Of the strains that have been compared to the sequenced MSB8, RQ2, PB1platt and S1-L12B share the highest level of genome conservation with MSB8. Only 129 ORFs in the MSB8 genome (1866 ORFs in total) did not have homologues in the RQ2 genome. These include 45 hypothetical proteins and 13 conserved hypothetical proteins, as well as 23 (18% of total that are absent) that are involved in transport. Of these 129, 18 occur as single ORFs, and the remaining correspond to islands that range in size from 2 kb to 38 kb. For strain S1-L12B, 9.4 % of the ORFs in MSB8 do not have homologs in this genome. Of these 174, 48 occur as single ORFs, and there are a total of 22 islands larger than 2 kb that are absent. Sixty-six ORFs correspond to hypothetical proteins, and 29 ORFs correspond to conserved hypothetical proteins. In addition, 6.9% are devoted to transport. Ten percent (186) of the MSB8 ORFs do not have homologs in PB1 (55 hypothetical proteins, 33 conserved hypotheticals), 16% of which are involved in transport. There are a total of 18 islands greater than 2kb in size that are absent from this strain. *T. thermarum* LA3 appears to be the most distantly related to MSB8. Initial data analysis suggests that lateral gene transfer across hyperthermophiles may be mediated by repetitive sequences that can be found in all these species. Interestingly, there is a high percentage of genes that are shared between *T. maritima* MSB8 and *Thermotoga* strain PB1platt that was isolated from an oil field in Alaska.

We have also designed, ordered and received primers to create a *Pyrococcus furiosus* genome microarray, and are in the process of diluting the primers and generating the PCR products that represent the entire genome. It is anticipated that we will conduct experiments similar to the *Thermotoga* comparative genome hybridization study.

B29

The Dynamics of Cellular Stress Responses in *Deinococcus radiodurans*

Michael J. Daly[1], Jizhong Zhou[2], James K. Fredrickson[3], Richard D. Smith[3], Mary S. Lipton[3], and Eugene Koonin[4]

[1]Uniformed Services, University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814; Tel: 301-295-3750; [2]Oak Ridge National Laboratory, Oak Ridge, TN; [3]Pacific Northwest National Laboratory, Richland, WA; and [4]National Center for Biotechnology Information, NIH, Bethesda, MD

Deinococcus radiodurans (DEIRA) is the most characterized member of the radiation resistant bacterial family *Deinococcaceae*. It is non-pathogenic, amenable to genetic engineering, and historically best known for its extreme resistance to gamma radiation [1]. The bacterium can grow and functionally express cloned foreign genes in the presence of 60 Gy /hour, and can survive acute exposures that exceed 15,000 Gy without lethality [1, 2]. How this feat is accomplished is unknown, and a long-term goal of our GTL project is a detailed understanding of the molecular pathways underlying this phenotype. Based on its remarkable robustness, DEIRA is also being developed for bioremediation of radioactive mixed waste sites containing radionuclides, heavy metals, and toxic organic compounds [2]. Using a combination of computational and whole-cell technologies, we are analyzing expression networks in DEIRA to map its cellular repair pathways, and also using information gleaned from this work to facilitate its development for bioremediation.

Whole genome sequencing, annotation, and comparative analyses for DEIRA [1, 3] have given rise to the development of new experimental whole-cell technologies dedicated to this organism. In collaboration with co-investigators at PNNL, our groups have developed proteomic methodology that uses high-resolution liquid chromatography and Fourier transform ion cyclotron resonance mass spectrometry to characterize an organism's dynamic proteome. Using this technology, >61% of the predicted proteome of DEIRA has been characterized with high confidence [4]. This represents the broadest proteome coverage for any organism to date. And, in collaboration with co-investigators at ORNL and NCBI, we have constructed a whole-genome microarray (WGM) for DEIRA and used it to examine global RNA expression dynamics during recovery from high-dose irradiation [5].

Already, proteomic and WGM research has revealed an unprecedented view of the molecular systems involved in the resistance phenotypes of DEIRA, and has also facilitated the construction of DEIRA strains capable of detoxifying highly radioactive waste environments. For example, with respect to novel genes that may be involved in radiation resistance, we have confirmed the involvement of several that show marked induction following irradiation [5]. This work has also alerted us to how metabolic strategies, not generally associated with DNA protection or repair, could enhance its resistance functions. DEIRA switches its metabolic pathways in response to irradiation, minimizing oxidative stress production and optimizing recovery [6]. We now believe that a comprehensive knowledge of DIERA metabolism is key to advancing our understanding of its extreme radiation resistance, as well as extending its intrinsic metabolic functions for bioremedia- tion. In the past, our goal of engineering DEIRA for complete mineralization of toluene could not be reached because of uncertainties regarding engineering strategies relating to its metabolic configuration. These uncertainties were overcome following proteomic and WGM analyses that untangled some of the complexities of DEIRA metabolism, and revealed how DEIRA intermediary metabolism could be integrated with complete toluene oxidation. As a result, we have successfully engineered DEIRA strains that can completely mineralize toluene, as demonstrated in natural sediment and groundwater analogs of DOE contaminated environments [7]. In summary, the GTL program is providing us a unique opportunity to bring together state-of-the-science high-throughput whole-cell technologies to explore fundamental and applied aspects of *Deinococcus radiodurans*.

1. K. S. Makarova, *et al.* (2001) The genome of the extremely radiation resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomics. *Microbiology and Molecular Biology Reviews* **65**, 44–79.

2. H. Brim, *et al.* (2000) Engineering *Deinococcus radiodurans* for metal remediation in radioactive mixed waste environments. *Nature Biotechmology* **18**, 85–90.

3. O. White, *et al.* (1999) Sequencing and functional analysis of the *Deinococcus radiodurans* genome. *Science* **286**, 1571–1577.

4. M. S. Lipton, *et al.* (2002) Global analysis of the *Deinococcus radiodurans* R1 proteome using accurate mass tags. *Proc. Natl. Acad. Sci. USA*, **99**, 11049–11054.

5. Y. Liu, J. Zhou, A. Beliaev, J. Stair, L. Wu, D.K. Thompson, D. Xu, A. Venkateswaran, M. Omelehenko, M. Zhai, E. K. Gaidamakova, K. S. Makarova, E. Koonin, and M. J. Daly (2002) Transcriptome dynamics of *Deinococcus radiodurans* recovering from ionizing radiation. Submitted.

6. A.Vasilenko, A.Venkateswaran, H. Brim, Y. Liu, J. Zhou, K. S. Makarova, M. Omelchenko, D. Ghosal, and M. J. Daly (2002) Relationship between metabolism, oxidative stress and radiation resistance in the family *Deinococcaceae*. Submitted.

7. H. Brim, J. P. Osborne, A Venkateswaran, M. Zhai, J. K. Fredrickson, L. P. Wackett, and M. J. Daly (2002). Facilitated Cr(VI) Reduction by *Deinococcus radiodurans* Engineered for Complete Toluene Mineralization. Submitted.

B9

Uncovering the Regulatory Networks Associated with Ionizing Radiation-Induced Gene Expression in *D. radiodurans* R1

John R. Battista[1] (jbattis@lsu.edu), Ashlee M. Earl[1], Heather A. Howell[2], and Scott N. Peterson[2]

[1]Department of Biological Sciences, Louisiana State University and A & M College, Baton Rouge, LA 70803; and [2]The Institute for Genomic Research, Rockville, MD 20850

In an effort to determine which genes are LexA regulated in *D. radiodurans* a lexA defective strain of *D. radiodurans* R1, GY10912, was evaluated using microarray analysis. Under normal, unstressed conditions over 100 transcripts were more abundant in GY10912 than in the wild-type strain suggesting that a large fraction (3%) of *D. radiodurans* genome is regulated by this repressor. However, only 22 genes from the LexA controlled gene set overlapped with the 71 genes previously determined to be stress induced following exposure to ionizing radiation in wild type R1. There is absolutely no overlap between the 'classical' SOS regulon of *E. coli* and LexA controlled genes in *D. radiodurans*. When a 3,000Gy dose of ionizing radiation is administered to GY10912 only 7 additional genes are induced including recA. Since a LexA defect does not render *D. radiodurans* sensitive to ionizing radiation, it is assumed that the cell only needs to up-regulate these 29 loci: the 22 LexA dependent loci and the 7 LexA independent loci. The LexA independent induction is in part controlled by the IrrE protein (DR0167). IrrE is a positive effector that when inactivated results in loss of ionizing radiation resistance. Loss of IrrE prevents the expression of 13 loci that are normally induced in response to ionizing radiation. All 13 of these loci overlap with those genes thought to confer resistance to GY10912.

B31

Analysis of Proteins Encoded on the *S. oneidensis* MR-1 Chromosome, Their Metabolic Associations, and Paralogous Relationships

Margrethe H. Serres*, Maria C. Murray, and **Monica Riley**

*Presenter

Marine Biological Laboratory, Woods Hole, MA 02543

Proteins encoded by the *Shewanella oneidensis* MR-1 chromosome have been analyzed for their sequence similarity to proteins encoded in 49 completely sequenced microbial genomes. It is our goal to elucidate the metabolic pathways in *S. oneidensis* in order better to understand the metabolic capabilities of this versatile organism. The genome is also being analyzed for paralogous or sequence similar proteins within the chromosome. Such protein families have been found useful in assigning putative functions to gene products and have provided insight into the evolution of the genome and the functions encoded within.

Sequence similarity searches were done using Darwin and an alignment requirement of at least 83 amino acids. Sequence matches were found for 82% of the encoded proteins at a similarity of <=250 PAM units. *Vibrio cholerae*, *Pseudomonas aeruginosa*, and *Yersinia pestis* showed the highest percentage of best hits at 23%, 10%, and 7%, respectively.

To identify metabolic pathways we conferred with the GenProtEC and EcoCyc/BioCyc databases and with published literature. We used the MultiFun classification system consisting of the following classes; metabolism, information transfer, regulation, transport, cell processes, cell structure, location, extra-chromosomal origin, DNA sites and cryptic genes. Currently more than 3800 cell function assignments have been made to 1560 *S. oneidensis* proteins. Among these, 740 proteins were assigned to metabolism, including 193 proteins classified as having a role in respiration. An

overview of pathways, missing steps and similarity to other genomes will be presented.

In order to determine paralogous protein families, proteins encoded by fused or composite genes have to be identified and their sequences separated into entities (modules) of independent evolutionary origin. The pair-wise alignments from the Darwin analysis described above have been used to discern modules. Currently 170 (4%) of the chromosomally encoded *S. oneidensis* proteins have been identified as composite. This number is slightly higher than for the *E. coli* genome. An initial grouping of protein modules is in progress and will be presented.

B33

Finishing and Analysis of the *Nostoc punctiforme* Genome

S. Malfatti[1] (malfatti3@llnl.gov), L. Vergez[1], N. Doggett[2], J. Longmire[2], R. Atlas[3], J. Elhai[4], J. Meeks[5], and **P. Chain**[1]

[1]Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA; [2]Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM; [3]Department of Biology, University of Louisville, Louisville, KY; [4]Department of Biology, University of Missouri, St. Louis, MO; and [5]Section of Microbiology, University of California, Davis, CA

In an effort to explore the role of diverse microorganisms in global carbon sequestration, the DOE's JGI has drafted the genomic sequence of several microorganisms that play unique roles in soil and ocean ecosystems. Elucidation of their genetic content will allow de novo identification of metabolic pathways relevant to understanding the physiological and genetic controls of photosynthesis, nitrogen fixation and general carbon cycling. A major contributor to the sequestration of CO_2 in organic compounds, *Nostoc punctiforme* is a unique nitrogen-fixing cyanobacterium that can differentiate into three types of specialized cells (N_2-fixing, motile and spore-like), is capable of forming symbiotic associations with fungi and plants and has the ability to grow rapidly under completely dark heterotrophic conditions. We have undertook the task of finishing and analyzing the genome of *N. punctiforme* strain ATCC 29133, which is likely to yield novel information on global regulation of multiple developmental pathways, symbiotic association, and phylogenetic relation to the other cyanobacteria including the now complete *Prochlorococcus* and *Synechococcus* strains.

The genome of *N. punctiforme* is very large for a prokaryote, nearly 10Mb, which makes this an attractive genome to complete. There are currently near 100 prokaryotic genomes that have been sequenced and closed, yet the average finished genome size is only ~3Mb and fewer than 10 have a genome size larger than 6Mb. Thus, there is quite possibly a bias in our prokaryotic genome knowledge toward smaller genomes and there may be some distinctive features in larger prokaryotic genomes that require finer resolution.

The genome was drafted by the JGI to 10-fold coverage due to the large number of contigs observed at 5-, 6- and 7-fold; however, the number of contigs greater than 2kb with >10 reads remained essentially the same, near 230 contigs. Additional rounds of automated primer design for finishing purposes has not significantly reduced this number. The reason for the difficulties in finishing are likely due to the prevalence of repeated elements within the genome. This is supported by the high coverage (25- to 50-fold) of entire contigs of large size (up to 110 kb). The identification and resolution of these repeated elements is being performed with the use of a fosmid scaffold along with a suite of computational tools. An overview of the progress in finishing *N. punctiforme* will be presented.

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.

B35

In Search of Diversity: Understanding How Post-Genomic Diversity is Introduced to the Proteome

Barry Moore, Chad Nelson, Norma Wills, John Atkins, and **Raymond Gesteland** (ray.gesteland@genetics.utah.edu)

Department of Human Genetics, University of Utah, Salt Lake City, UT

Analysis of an organism's genome is the primary tool for understanding the diversity of protein products present in an organism. However, as proteomics focuses more work on the proteins themselves, it is apparent that the diversity of the

proteome goes well beyond that indicted in the genome. Much of the energy and resources in proteome projects have been invested in various peptide based mapping strategies—principally by 2-D gel electrophoresis followed by MALDI-MS analysis. One shortcoming of peptide approaches, however, is that they do not allow for determination of the full-length protein mass. This means that while it can be determined that a particular region of the genome (the ORF) is translated into protein, no direct evidence is provided about the full-length protein product, and there is no indication of the diversity of different proteins products translated from that ORF.

A number of studies have indicated that regions of sequenced genomes not currently identified as ORFs are translated, or are likely to be translated. Furthermore, numerous mechanisms are known by which diversity is introduced to an organism's proteome post-genomically, and there are a wide variety of bioinformatics algorithms either available or in development that attempt to predict these events from genome sequence. Ultimately, we must go to the proteins themselves to fully understand when and where post genomic diversity is arising and to refine our ability to predict from genome analyses the diversity of proteins derived from an organism's genome. We are developing a dual affinity tagging system in the yeast *Saccharomyces cerevisiae* that allows us to build fusions to regions of the genome where novel ORFs are predicted, or where non-standard translational events are expected to occur. Fusion proteins are expressed in yeast and purified by affinity chromatography to produce a protein pure enough for analysis by electrospray mass spectrometry. Analysis of a highly accurate full length mass in combination with tryptic peptide fragments and MS/MS analysis of those fragments allows a thorough understanding of the proteins full length covalent state, and the diversity of products produced from a particular ORF.

B37

The Microbial Proteome Project: A Database of Microbial Protein Expression in the Context of Genome Analysis

Carol S. Giometti[1] (csgiometti@anl.gov), Gyorgy Babnigg[1], Sandra L. Tollaksen[1], Tripti Khare[1], Derek R. Lovley[2], James K. Fredrickson[3], Kenneth H. Nealson[4], Claudia I. Reich[5], Gary J. Olsen[5], Michael W. W. Adams[6], and John R. Yates III[7]

[1]Argonne National Laboratory; [2]University of Massachusetts; [3]Pacific Northwest National Laboratory; [4]University of Southern California; [5]University of Illinois at Urbana; [6]University of Georgia; and [7]The Scripps Research Institute

Although complete genome sequences can be used to predict the proteins expressed by a cell, such predictions do not provide an accurate assessment of the relative abundance of proteins under different environmental conditions. In addition, genome sequences do not define the subcellular location, biomolecular and cofactor interactions, or covalent modifications of proteins that are critical to their function. Analysis of the protein components actually produced by cells (i.e., the proteome) in the context of genome sequence is, therefore, essential to understanding the regulation of protein expression. As the number of complete microbial genome sequences increases, vast amounts of genome and proteome information are being generated.

In parallel with the proteome analysis of numerous microbial systems, researchers at Argonne National Laboratory are developing methods for managing and interfacing the diverse data types generated by both genome and proteome studies as part of Argonne's Microbial Proteome Project. The goal is to provide users with a highly interactive database that contains proteome information in the context of genome sequence in formats that are conducive to data interrogations that will provide answers to biological questions. To achieve that goal, three World Wide Web-based databases are currently being developed and maintained: Proteomes2, ProteomeWeb, and GelBank.

The Proteomes2 database (proteomes2.bio.anl.gov) serves as a Laboratory Information Management System for the management of sample data and related two-dimensional gel electrophoresis (2DE) patterns. This password-protected site provides DOE project collaborators with access to

data access from multiple sites through the Internet. The database currently contains the experimental details for approximately 1000 samples from seven different microbes (*Shewanella oneidensis, Geobacter sulfurreducens, Prochlorococcus marinus, Methanococcus jannaschii, Pyrococcus furiosus, Rhodopseudomonas palustris,* and *Deinococcus radiodurans*) and links each sample with multiple protein patterns. Over 4000 protein pattern images are currently accessible.

ProteomeWeb (http://ProteomeWeb.anl.gov) is an interactive public site that provides the identification of expressed microbial proteins, links to genome sequence information, tools for mining the proteome data, and links to metabolic pathways. Data from proteome analysis experiments are included in the ProteomeWeb database when genome sequences are deposited in GenBank. Currently, the results from experiments designed to alter protein expression in *M. jannaschii* are accessible on this site, and results from *S. oneidensis* experiments are in the process of being incorporated.

GelBank currently includes the complete genome sequences of approximately 90 microbes and is designed to allow queries of proteome information. The database is currently populated with protein expression patterns from the Argonne Microbial Proteomics studies and will accept data input from outside users interested in sharing and comparing proteome experimental results.

Oracle9i RDBMS — the common foundation for all three Argonne proteomics databases — allows the integration of genome sequences, sample descriptors, protein expression data (e.g., 2DE images and numerical data), peptide masses, and annotation. The database management architecture being developed currently addresses protein expression data from 2DE analysis of protein mixtures, but it is being designed to have the flexibility to accommodate mass spectrometry data as well.

This research is funded by the United States Department of Energy, Office of Biological and Environmental Research, under Contract No. W-31-109-ENG-38.

B39

Analysis of the *Shewanella oneidensis* Proteome in Cells Grown in Continuous Culture

Carol S. Giometti[1] (csgiometti@anl.gov), **Mary S. Lipton**[2] (mary.lipton@pnl.gov), Gyorgy Babnigg[1], Sandra L. Tollaksen[1], Tripti Khare[1], James K. Fredrickson[2], Richard D. Smith[2], Yuri A. Gorby[2], and John R. Yates III[3]

[1]Argonne National Laboratory; [2]Pacific Northwest National Laboratory; and [3]The Scripps Institute

We are using two complementary methods to identify and quantify the proteins expressed by *Shewanella oneidensis* MR-1 grown under different conditions in continuous culture. Cells are grown in chemostats under aerobic, O_2-limited, or anaerobic conditions with fumarate provided as the electron acceptor. Cells are harvested and aliquots of the same samples are analyzed for protein by using (1) two-dimensional gel electrophoresis (2DE) coupled with peptide mass analysis and (2) capillary liquid chromatography separations coupled with Fourier transform ion cyclotron resonance (FTICR) mass spectrometry. The 2DE patterns readily reveal statistically significant differences in protein abundance and in protein post-translational modifications that result from different growth conditions. The proteins that differ significantly in expression are then identified on the basis of their tryptic peptide masses. In parallel with the 2DE analysis, we are using accurate mass tags (AMTs) to identify all of the *S. oneidensis* proteins expressed under specific growth conditions and to detect quantitative differences by using stable isotope labeling methods. By using these two different protein separation and detection methods, we are able to more comprehensively analyze the proteome than by using either method alone. Specifically, 2DE provides a rapid turnaround "snap shot" of the major protein differences (including post-translational modifications) under different growth conditions, and AMTs provide a comprehensive inventory of the expressed proteins in each sample.

This work is funded by the U.S. Department of Energy, Office of Biological and Environmental Research, under contract No. W31-109-ENG-38 (Argonne National Laboratory) and contract No. DE-AC06-76RLO1830 (Pacific Northwest National Laboratory).

B41

The Molecular Basis for Metabolic and Energetic Diversity

Timothy Donohue[1] (tdonohue@bact.wisc.edu), Jeremy Edwards[2], Mark Gomelsky[3], Jonathan Hosler[4], Samuel Kaplan[5], and William Margolin[5]

[1]Bacteriology Department, University of Wisconsin-Madison; [2]Chemical Engineering Department, University of Delaware; [3]Department of Molecular Biology, University of Wyoming; [4]Department of Biochemistry, University of Mississippi Medical Center; and [5]Department of Microbiology & Medical Genetics, University of Texas Medical School at Houston

Our long-term goal is to engineer microbial cells with enhanced metabolic capabilities. As a first step, this team of scientists and engineers seeks to acquire a thorough understanding of energy-generating processes and genetic regulatory networks of the photosynthetic bacterium, *Rhodobacter sphaeroides*. The ability to capitalize on the metabolic activities of this versatile bacterium was increased by the completion of the *R. sphaeroides* genome sequence at the DOE-supported Joint Genome Institute. The *R. sphaeroides* Genomes to Life Consortium is deciphering important energy-generating activities of this bacterium and studying the assembly and operation of energy generating machines. The long term goals of these efforts are to acquire the information needed to design microbial machines that degrade toxic compounds, remove greenhouse gases, or synthesize biodegradable polymers with increased efficiency. At the February, 2003 workshop, we will provide a progress report on our analysis of the metabolic capabilities of this facultative microorganism, particularly on the identification of proteins and regulators that are central to growth via respiration and the utilization of solar energy by photosynthesis.

In one line of experiments, we have begun to characterize the multitude of aerobic respiratory enzymes that this bacterium uses to generate energy in the presence of O_2. The goal of these experiments is to engineer strains that can more efficiently extract energy from nutrients in the presence of O_2.

We have begun to visualize the assembly of the photosynthetic apparatus that this bacterium uses to harvest solar energy. When this analysis is combined with modeling of solar energy utilization by this well-studied microbe, we hope to identify principles that will allow us to engineer microbes with increased ability to generate energy from light.

The feasibility of generating strains with increased capacity for using solar radiation is enhanced by our identification of previously unrecognized pigment-binding proteins in the photosynthetic apparatus of *R. sphaeroides*. Efforts are currently underway to understand the contribution of these new pigment-binding proteins to the utilization of light energy that is critical for growth of photosynthetic bacteria and plants.

Simultaneously we have initiated a program to identify new regulators of the solar lifestyle of this bacterium and to use gene chips to identify additional proteins that could be critical to the utilization of light energy. Both of these approaches have provided new candidates that are currently being analyzed for their role as regulators or contributors to the utilization of solar energy.

We hope to illustrate why this cross-disciplinary, systems approach to the analysis of energy generation by this facultative bacterium can provide new insights into fundamental aspects of energy generation and the utilization of solar energy by this and other photosynthetic organisms.

B43

Integrative Studies of Carbon Generation and Utilization in the Cyanobacterium *Synechocystis sp.* PCC 6803

Wim Vermaas[1] (wim@asu.edu), Robert Roberson[1], Julian Whitelegge[2], Kym Faull[2], Ross Overbeek[3], and Svetlana Gerdes[3]

[1]Arizona State University; [2]University of California Los Angeles; and [3]Integrated Genomics, Inc.

The research focuses on photosynthetic and respiratory electron transport and on carbon utilization in the cyanobacterium *Synechocystis sp.* PCC 6803. The research compares wild type with targeted mutants lack the photosystems, the terminal oxidases, succinate dehydrogenase, and/or other complexes that are important for photosynthesis and respiration.

Physiological Analysis

1. Carbon utilization: Mutants lacking terminal oxidases were found to accumulate large inclusions that resembled poly-β-hydroxybutyrate bodies. As these inclusions were by far most prevalent in mutants lacking terminal oxidases, they most likely result from fermentation reactions. The very significant accumulation of polyhydroxyalkanoates (up to about 25% of the total cell volume) in *Synechocystis* mutants lacking the terminal oxidases suggests a potential application of this strain in light-driven production of "bioplastic" materials.

2. Membrane biogenesis: One major question that remains in chloroplasts and cyanobacteria is how thylakoids are formed. A gene that apparently is involved with thylakoid biogenesis has been interrupted at various positions, and the interruption that has segregated (with the site of insertion close to the end of the gene) showed a normal ultrastructural phenotype. We are also inactivating genes involved with cell division, in the hope of creating a larger cyanobacterial cell that is more easily analyzed by light microscopy techniques.

3. Toward a system suitable for crystal structure analysis: Synechocystis has proven to be an excellent model system from a standpoint of molecular genetics, physiology, etc. However, for elucidation of structures of membrane proteins (another goal that should be pursued in microbial cell projects) thermophilic cyanobacteria, such as *Thermosynechococcus elongatus*, for which a genome sequence is known, are very much preferable. For full utility of this system, mutants that possibly impair the photosystems should be used and therefore the organism should be able to grow photoheterotrophically, which it currently is unable to do. Therefore, we have initiated random mutagenesis experiments to see whether a *Thermosynechococcus* mutant can be obtained that can grow under conditions other than photoautotrophic ones.

Structural Analysis

Much progress has been made in ultrastructural preservation and documentation of cytoplasmic organization in *Synechocystis sp.* PCC 6803.

1. Three-dimensional reconstruction: Using thick (0.2-0.3 μm) sections of *Synechocystis*, we have collected data on high- and intermediate-voltage transmission electron microscopes (0.2-1 MeV) to three-dimensionally image the intracellular organization of *Synechocystis sp.* PCC 6803 cells. In this procedure, thick sections are incrementally tilted from -60^{o} to +60^{o} and serial tilt views electronically captured at 1.5^{o} intervals. We have produced the first high-resolution three-dimensional images of a cyanobacterium, and are now in the process to identify physical relationships between thylakoid and cytoplasmic membranes, the fate of membranes upon cell division, etc. Clearly, tomography using thick sections is an extremely powerful technique that is providing new insights into the ultrastructural organization of the cell.

2. Membrane organization: We observe "thylakoid centers" in dividing cells. These structures were discovered about two decades ago in another cyanobacterium, and apparently have been forgotten since. Thylakoid centers are located at the point of apparent origin of thylakoid membranes (i.e., the point of thylakoid membrane convergence) near the cytoplasmic membrane. At these specialized regions, membranes either terminate or turn 180° in close proximity to the cytoplasmic membrane. The thylakoid center apparently extends as a tube-like structure for at least 200 nm. We are interested in exploring the role and composition of these thylakoid centers.

Proteomics

With the complete genome sequence available, comparative proteome analysis is underway with the goal of providing data that will integrate with the ultrastructural work and in mutants of photosynthesis and respiration. By using sub-fractionation of soluble or membrane preparations it is possible to preserve native protein complexes providing functional insights. Soluble fractions are examined by size-exclusion chromatography and 2D-gel electrophoresis for evidence of the carboxysome seen in EM work to be more abundant in the SDH-less mutant. Membrane preparations are sub-fractionated by sucrose-gradient centrifugation to separate thylakoids from cytoplasmic membranes, and to search for preparations that may be enriched in thylakoid centers, in order to better understand the specialization of each membrane system and the trafficking between them. Membrane protein complexes are being separated by size-exclusion chromatography under non-denaturing conditions after solubilization of membranes with detergents as the first dimension of a 2D chromatography system. A second dimension denaturing chromatography system incorporates intact protein mass measurement (LC-MS+). Fractions collected during LC-MS+ are used for protein identification

by fingerprinting and sequencing and are also available for further analysis by ultra-high resolution Fourier transform mass spectrometry (FT-MS). The successful analysis of intact photosystem I reaction-center polypeptides PsaB and PsaA of mass 81,167 and 82,876 Da, respectively, demonstrates that *intact mass proteomics* can be applied to large integral membrane proteins with resolution sufficient to detect single methionine oxidations at this size.

Bioinformatics

Integrated Genomics began work on the GTL Project in October 2002; hence progress in this area has been somewhat limited (but not insignificant) so far. The initial objectives are:

1. To support the wet lab characterization of critical genes by making specific predictions that will be tested experimentally by our partners.

2. To produce a detailed metabolic reconstruction to support a more complete understanding and modeling of this organism.

Integrated Genomics is coordinating closely with other members of the group to establish priorities relating to prediction of gene function. During the first month, we have made two predictions, one addressing the lycopene cyclase, and one related to a gene associated with phycobilisomes. The lycopene cyclase is by far the more interesting of these predictions, and other members of the group have begun experiments to confirm or reject our candidate gene. We have now begun a tabulation of predictions.

The second broad goal is to develop a detailed metabolic reconstruction for *Synechocystis sp.* PCC 6803. In particular, two aspects of this effort are pursued: (1) development of a detailed graphical/web-based interface to a reconstruction connecting functional components to genes in the organism, and (2) development of a precisely encoded version of the reaction network in a form suitable for supporting modeling efforts.

Altogether, the GTL project on *Synechocystis sp.* PCC 6803 provides integrated functional-genomics information regarding the structure of the organism and the function of photosynthesis and respiration related processes, based on physiological, ultrastructural, proteomics, and bioinformatics experimentation on the wild type and targeted mutants.

Technology Development

B45

Comparative Optical Mapping: A New Approach for Microbial Comparative Genomics

Shiguo Zhou[1] (szhou@lmcg.wisc.edu), Thomas S. Anantharaman[2], Erika Kvikstad[1], Andrew Kile[1], Mike Bechner[1], Wen Deng[3], Jun Wei[3], Valerie Burland[3], Frederick R. Blattner[3], Chris Mackenzie[6], Timothy Donohue[4], Samuel Kaplan[6], and **David C. Schwartz**[1,5] (dcschwartz @facstaff.wisc.edu)

[1]Laboratory for Molecular and Computational Genomics, [2]Animal Health and Biomedical Sciences, [3]Laboratory of Genetics, [4]Department of Bacteriology, and [5]Department of Chemistry, University of Wisconsin-Madison, Madison, WI 53706; and [6]Department of Microbiology and Molecular Genetics, University of Texas-Houston Medical School, 6431 Fannin St., Houston, TX 77030

The recent plethora of sequenced genomes has just ushered in a new era of genetics-based research. Although an impressive number of species have been, or are planned to be sequenced, the full value of such efforts will be fully accrued when patterns of variation can be discerned and annotated for many strains or isolates within a given species. As such, bacteria are an ideal place to start investigations aimed at the discernment of genome rearrangements, chromosome deletions, and horizontal transfer of foreign DNA, since these events help drive bacterial evolution. For example, genome remodeling events may cause irreversible gene loss, or add novel functionalities to an organism. Unfortunately, current approaches do not adequately identify and characterize such large-scale genomic rearrangements. The Optical Mapping System, developed in our laboratory creates high resolution maps of entire genomes, using DNA directly extracted from cells—this approach obviates the need for libraries, PCR, and probe technologies. The system uses a complex blend of single molecule technologies to enable high throughput and the construction of reliable maps. This capability has been proven by the construction and sequence comparison of 13 bacterial optical maps, and 3 parasites. Recent advances in both throughput and resolution of the Optical Mapping System has enabled genomic comparisons amongst different strains of the same species or closely related species, allowing for the pinpoint discernment of insertions, deletions and rearrangements. Comparisons of optical maps vs. in silico maps, and in silico vs. in silico maps constructed for two strains of *Yersinia pestis* (CO-92 biovar Orientalis and KIM), *E. coli* K12 and *Shigella flexneri* 2a, two strains of *S. flexneri* (2a and Y), and two strains of *Rhodobacter sphaeroides* (2.4.1 and ATCC 17029) have revealed regions of homology, insertion sites and a panoply of rearrangements. These results portend the wide use of Optical Mapping to uniquely provide genome structural details for a large number of strains, isolates or even closely related species, in ways that would complement direct sequence analysis.

B47

Optical Mapping of Multiple Microbial Genomes

Shiguo Zhou[1] (szhou@lmcg.wisc.edu), Michael Bechner[1], Erika Kvikstad[1], Andrew Kile[1], Susan Reslewic[1], Aaron Anderson[1], Rod Runnheim[1], Jessica Severin[1], Dan Forrest[1], Chris Churas[1], Casey Lamers[1], Samuel Kaplan[4], Chris Mackenzie[4], Timothy J. Donohue[2], and **David C. Schwartz**[1,3] (dcschwartz @facstaff.wisc.edu)

[1]Laboratory for Molecular and Computational Genomics, [2]Department of Bacteriology, and [3]Department of Chemistry, University of Wisconsin-Madison, Madison, WI 53706; and [4]Department of Microbiology and Molecular Genetics, University of Texas-Houston Medical School, 6431 Fannin St., Houston, TX 77030

Our laboratory has developed Optical Mapping, which is a proven system for the construction of ordered restriction maps from individual DNA molecules directly extracted from cells. Such maps have utility in large-scale sequencing efforts by providing scaffolds for assembly, and as an independent means for validation, since entire genomes are mapped without the use of clones or PCR amplicons. Given the major increase in the throughput of Optical Mapping, we have used this system to map multiple microbial genomes,

which included; *Thalassiosira pseudonanna* (diatom), *Enterococcus faecium*, *Pseudomonas fluorescens*, *Rhodobacter sphaeroides* and *Rhodospirillum rubrum*. These efforts were funded by DOE to help expedite and validate parallel sequencing projects. Our optical mapping data showed that *T. pseudonanna* has a haploid genome size of 33.8 Mb, possessing 22 chromosomes ranging from 658 kb to 3,322 kb, while other genomes had the following features: *E. faecium* (2.8 Mb), *P. fluorescens* (6.9 Mb, which is 1.4 Mb larger than expected size of 5.5 Mb), *R. rubrum* (4.2 Mb instead of the expected 3.4 Mb), and *R. sphaeroides* (4.2 Mb). Overall the map resolution varied from a low resolution map of average restriction fragment size of 50 kb for *R. rubrum* (*Xba*I), to a high resolution map with an average fragment size of 6.0 kb for the *R. sphaeroides* *Hind*III map. Comparison of optical maps of *R. sphaeroides* with the nascent sequence contigs of the genome sequencing project detailed the utility of optical maps in expediting sequencing projects by the identification of misassemblies within nascent sequence contigs, gap characteristics and the orientation of sequence contigs.

B49

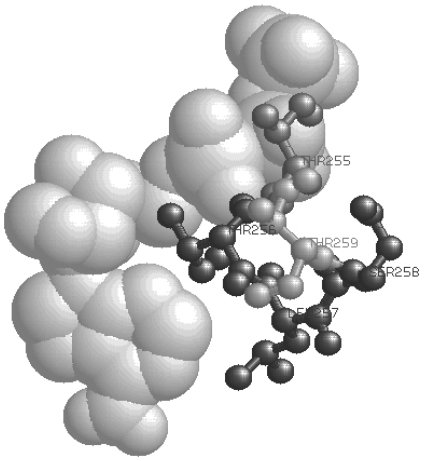
Identification of ATP Binding Proteins within Sequenced Bacterial Genomes Utilizing Phage Display Technology

Suneeta Mandava, Lee Makowski, and **Diane J. Rodi** (drodi@anl.gov)

Combinatorial Biology Unit, Biosciences Division, Argonne National Laboratory, Argonne, IL 60439

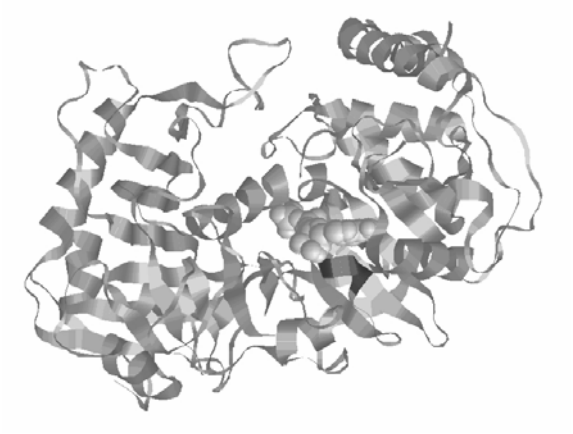
This project is applying a novel approach to genome-wide identification of small molecule binding proteins. Our ability to rapidly sequence prokaryotic genomes is generating an unprecedented amount of DNA sequence data. In spite of the large number of functional genomics tools currently available, typically about 40% of predicted ORFs remain unidentified in terms of function. Our results demonstrate that the similarity between the sequence of a protein and the sequences of phage-displayed peptides affinity-selected against small molecules can be predictive for that protein binding to the small molecule. In this project, we utilize tagged derivatives of the common metabolite ATP fixed to a solid surface as bait to capture peptide-bearing phage particles from solution. Population analysis of hundreds of these captured peptides demonstrates that subpopulations represent portions of known ATP-binding motifs.

To test our ability to predict ATP binding site locations within a protein the similarity between the sequences of our affinity selected ATP peptides and the sequences of known ATP binding proteins from the PDB were calculated. Shown here is an example of this technique applied to the ATP-binding protein phosphoenolpyruvate carboxykinase: (Renderings were carried out using RASMOL in which the segments of maximum similarity are designated in red, and blue the lowest similarity.)



To determine whether or not this method can be used as a global approach to predicting which proteins bind ATP within a sequenced genome, as well as to identify the position of those ATP binding sites, we have developed software which compares the ATP-binding peptide pool to entire genome protein sequences. Comparison of the ATP-binding peptide populations for affinity to two sets of data, ATP-binding proteins in the Protein Data Bank and the entire *E. coli* K12 proteome confirmed that annotation of open reading frames for small molecule binding is possible using this method. Successful identification of residues within 7 Å of the ATP ligand within PDB structures is accomplished in 75% of proteins tested. Alignment of all the proteins in the *E. coli* proteome by peptide-similarity score segregates ATP binders to the top of the list when compared to control scores obtained with alternate ligand-selected pools of peptides.

The best characterized of the conserved sequence motifs that bind ATP or GTP is a glycine-rich region called a P-loop or Walker A box, which



typically forms a flexible loop between a beta-strand and an alpha-helix. In general, this loop has been shown to interact with one of the phosphate groups of the nucleotide within crystal structures. However, in two recently published P-loop-containing crystal structures, the P-loop motif is located far from the ATP-binding pocket of the protein. In these two cases maximum similarity with our ATP-binding peptide population resides within the actual ATP-binding site as opposed to the P-loop site. This discrepancy may be the result of crystal contacts that alter ATP binding, or a confirmation of the binding funnel theory of Nussinov, which describes the process of ligand binding as a series of short-lived conformational ensembles along a decreasing energy landscape. This interpretation of these observations predicts that although a P-loop motif is predictive of ATP binding, and is the initial site of an early-lived molecular handshake with ATP, the ATP ligand may not remain in the proximity of the P-loop in the final global free energy minimum conformation. This scenario similarly predicts that there may exist alternate primary sequence motifs predictive for ligand binding within the same protein sequence, such as the motif identified with our ATP-binding peptide pool.

Distribution of the software, analysis methods and data generated from this ongoing project is being accomplished by the construction of an ORACLE web-based database named RELIC (for REceptor LIgand Contacts). The architecture of the database is currently in place, and is scheduled to be made accessible to the public within the next three months, subsequent to optimization of software GUIs and a web-based users manual.

B51

Development of Vectors for Detecting Protein-Protein Interactions in Bacteria

Peter Agron and **Gary Andersen**
(andersen2@llnl.gov)

Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94551

We are interested in developing better approaches for mapping bacterial protein-protein interactions, particularly in *Caulobacter crescentus*, a model system for studying cellular differentiation and the cell cycle. Because of the advantages for high-throughput screening, our focus has been on using *Escherichia coli* as a host for two-hybrid assays. Initially, the BacterioMatch system from Stratagene was tested with 11 pairs of *Caulobacter* genes known to encode interacting proteins. Interactions were detected with several pairs, but the results were not found to be easily reproducible. Also, no interaction was observed with FtsZ, which is known to dimerize with high affinity. Therefore, we have constructed new vectors based on protein-fragment complementation with mouse dihydrofolate reductase (DHFR). In this assay, fusions to two portions of DHFR will reconstitute enzyme activity if tethered by protein-protein interactions, thus conferring trimethoprim resistance to the host as trimethoprim specifically targets the endogenous DHFR. The vectors have compatible replicons with different markers, thus allowing effective screening in *E. coli*. We are initially testing this

system with bacteriophage λ cI and Caulobacter *ftsZ*, two genes encoding proteins that dimerize with high affinity. Using this system, we are also placing the interacting domains of additional *C. crescentus* gene pairs in either the 5' or 3' orientation to each of the two portions of DHFR. A library of randomly sheared *C. crescentus* DNA fragments is being placed in either orientation to the carboxy-terminal DHFR protein fragments to screen for known interactions. Based on these results we will test up to 200 additional genes of interest for protein-protein interactions with the random-fragment *C. crescentus* library.

B53

Development and Use of Microarray-Based Integrated Genomic Technologies for Functional Analysis of Environmentally Important Microorganisms

Jizhong Zhou[1] (zhouj@ornl.gov), Liyou Wu[1], Xiudan Liu[1], Tingfen Yan[1], Yongqing Liu[1], Steve Brown[1], Matthew W. Fields[1], Dorothea K. Thompson[1], Dong Xu[1], Joel Klappenbach[2], James M. Tiedje[2], Caroline Harwood[3], Daniel Arp[4], and Michael Daly[5]

[1]Oak Ridge National Laboratory; [2]Michigan State University; [3]University of Iowa; [4]Oregon State University; and [5]Uniformed Services University of the Health Sciences

Microarrays constitute a powerful genomics technology for assessing whole-genome expression levels and defining regulatory networks. Under the support of the DOE Microbial Genome Program, whole genome microarrays containing individual open reading frames were constructed for *Shewanella oneidensis* MR-1 (~4.9 Mb), *Deinococcus radiodurans* (3.2 Mb), *Rhodopseudomonas palustris* (4.8 Mb), and *Nitrosomonas europaea* (2.7 Mb) at Oak Ridge National Laboratory. DNA fragments having less than 75% similarity to other sequences in the genome were selected as specific probes using our automatic primer design program, PRIMERGEN. The majority of the probes have the size of less than 1 kb. To obtain sufficient PCR products for array fabrication, genes were amplified 8 or 16 times in a total reaction volume of 100 ul. The amplified products were then pooled and purified using automated procedures. In total, approximately 4700, 3046, 4508, and 2354 genes were amplified from the *S. oneidensis*, *D. radiodurans*, *R. palustris*, and *N. europaea* genomes, respectively. Additional sets of primers were designed for genes that did not give expected amplification products or that gave low-quality amplification. A 50-mer specific oligonucleotide was designed and synthesized for genes that did not yield desired PCR products after two attempts with PCR primers. The genome coverage for all four bacteria ranged from 95 to 99%. Evaluation of microarray quality by direct scanning, PicoGreen staining and microarray hybridization indicated that the microarray printing quality was very good in terms of spot morphology, intensity and uniformity, and the constructed microarrays have been sent to many collaborators at different institutions.

To evaluate the performance of 50-mer oligonucleotide arrays for gene expression, 96 genes from *S. oneidensis* MR-1 were randomly selected. Microarrays containing various lengths of oligonucleotides (30-70 nt) and PCR products from the same set of genes were constructed. Preliminary hybridization results indicated that the sensitivity of oligonucleotide probes is significantly lower than that of PCR products. Further in-depth comparisons are ongoing.

Genetic mutants are important to functional genomics analysis. However, mutant generation is very time-consuming and labor-intensive. Thus, a simple, high-throughput single-strand oligonucleotide mutagenesis approach has been evaluated in *S. oneidensis*. A new genetic vector containing appropriate sets of genes was constructed. Our preliminary results indicated that the developed genetic vector can replicate well in *S. oneidensis* and confer desired properties. Further work includes the introduction of single-strand oligonucleotides for targeted gene deletion. Also, a protocol for efficient transformation of MR-1 cells by electroporation was developed, which is a critical step for developing a high-throughput single-stranded oligonucleotide-based genetic system.

B55

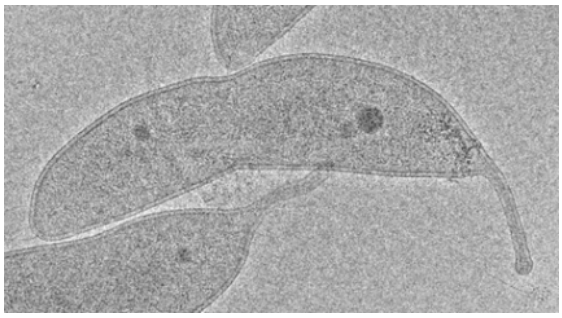
Electron Tomography of Whole Bacterial Cells

Ken Downing (khdowning@lbl.gov)

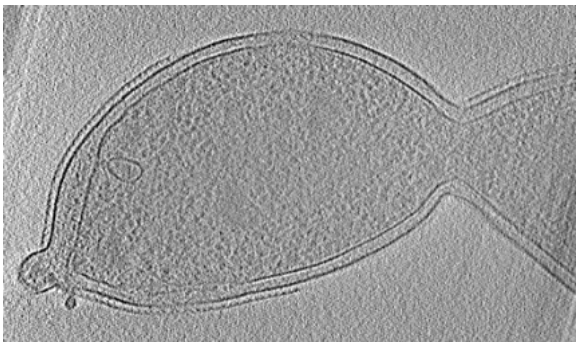
Lawrence Berkeley National Laboratory

In our initial work to develop electron tomography of intact cells and explore its limits of applicability, we have established culture and preparation conditions for a number of small microbes that may be potential targets for this work. These include *Magnetospirillum magnetotacticum*, *Caulobacter crescentis* and *Mesoplasma florum*. We have shown that we can record 2-D projection images by electron microscopy of each of these in frozen-hydrated preparations. We thus retain the native state with no stain or other contrast enhancements, but can see a wealth of internal structures. The mesoplasma is of particular interest since it is among the smallest and simplest living organisms, while these bacteria are sufficiently thin that we can obtain good data with an electron beam energy of 300-400 kV. We have collected several sets of preliminary tomographic data using facilities at the Max Planck Institute in Martinsried, Germany. 3-D reconstructions computed from these data are far more informative than the projection images. As expected, the 3-D maps show textures, representing distribution of proteins and/or nucleic acids, that vary within the organism, as well as some interesting and unexpected internal membrane structures. A number of steps need to be taken before we can begin to relate the densities seen in these reconstructions to individual protein complexes, but the preliminary data does suggest that this will indeed be feasible. Our own electron microscope, which will be especially well suited for this type of tomography, is presently being installed. Once it is in operation we will be able to further optimize our data recording protocols, including implementation of dual-axis tomograms, to improve the resolution and interpretability of the reconstructions.

Downing— Fig. 1. Electron micrograph of frozen-hydrated *Caulobacter crescentis*. The sample was rapidly frozen, with no stain or other contrast enhancement, preserving the native structure of the cell components.



Downing— Fig. 2. Section from a tomographic reconstruction of a cell undergoing division, with flagellum beginning to bud from left end. Patches of the periodic surface layer protein and internal membrane structures are visible in this section.



B57

Single Cell Proteomics—*D. radiodurans*

Norman J. Dovichi
(dovichi@chem.washington.edu)

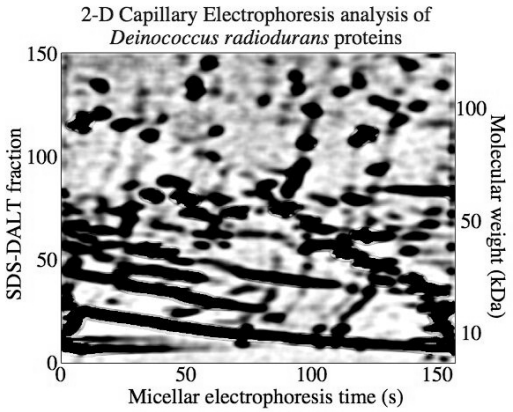
Department of Chemistry, University of Washington, Seattle, WA 98195-1700

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.

We have generated a number of fully automated two-dimensional capillary electrophoresis separations of proteins extracted from *D. radiodurans*. The figure below presents an example, in which the proteins from *D. radiodurans* are first subjected to capillary SDS-DALT separation, which is the capillary version of SDS-PAGE using replaceable polymers. Like SDS-PAGE, SDS-DALT separates proteins based on their molecular weight, with low molecular weight proteins migrating first from the capillary. Fractions are successively transferred to a second capillary, where proteins are separated in a sub-micellar electrophoresis buffer. Components are detected with an ultrasensitive laser-induced fluorescence detector at the exit of that capillary. Over 150 fractions are successively transferred from the first capillary to the second to generate a comprehensive analysis of the protein content of this bacterium. Data are stored in a computer and manipulated to form the pseudo-silver stain image of figure 1. We estimate that there are between 200 and 300 components resolved in this separation.

B59

Genomes to Proteomes to Life: Application of New Technologies for Comprehensive, Quantitative and High Throughput Microbial Proteomics

Richard D. Smith (rds@pnl.gov), James K. Fredrickson, Mary S. Lipton, David Camp, Gordon A. Anderson, Ljiljana Pasa-Tolic, Ronald J. Moore, Margie F. Romine, Yufeng Shen, Yuri A. Gorby, and Harold R. Udseth

Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99352

Achieving the Genomes to Life (GtL) Program goals will require obtaining a comprehensive systems-level understanding of the components and functions that give a cell life. At present our understanding of biological processes is substantially incomplete; e.g. we do not know with good confidence all the biomolecular players in even the most studied pathways and networks in microbial systems. It is clear that many important signal transduction proteins will be present only at very low levels (~ hundreds of copies per cell) and will provide extreme challenges for current characterization methods. There is also a growing recognition of the limitations associated with gene expression (e.g. cDNA array) measurements. Increasing evidence indicates that the correlation between gene expression and protein abundances can be low, and that the correlation between gene expression and gene function is even lower. Thus, global protein characterization (proteomic) studies actually complement gene expression measurements.

Successes in genome sequencing efforts have increased interest in proteomics and also provided an informatic foundation for high throughput measurements. As a result, a key capability envisioned for success of the GtL program is the ability to broadly identify large numbers of proteins and their modification states with high confidence, as well as to measure their abundances. The challenges associated with making useful comprehensive proteomic measurements include identifying and quantifying large sets of proteins that have relative abundances spanning more than six orders of magnitude, that vary broadly in chemical and physical properties, that have transient and low levels of modifications, and that are subject to endogenous proteolytic processing. Additionally, proteomic measurements should not

be significantly biased against e.g. membrane, large or small proteins. A related need is the ability to rapidly and reliably characterize protein interactions with other biomolecules, particularly their multi-protein complexes. The combined information on protein complexes and the changes observed from global proteome measurements in response to a variety of perturbations is essential for the development of detailed computational models for microbial systems and the eventual capability for predicting their response e.g. to environmental changes and mutations.

We report on development and application of new technologies for global proteome measurements that are orders of magnitude more sensitive and faster than existing technologies and that promise to meet many of the needs of the GtL program. The approaches are based upon the combination of nano-scale ultra-high pressure capillary liquid chromatography separations and high accuracy mass measurements using Fourier transform ion cyclotron resonance (FTICR) mass spectrometry. Combined, these techniques enable the use of highly specific peptide 'accurate mass and time' (AMT) tags. This new approach avoids the throughput limitations associated with other mass spectrometric technologies using tandem mass spectrometry (MS/MS), and thus enables fundamentally greater throughput and sensitivity for proteome measurements. Additional new developments have also significantly extended the dynamic range of measurements to approximately six orders of magnitude and are now providing the capability for proteomic studies from very small cell populations, and even single cells. A significant challenge for these studies is the immense quantities of data that must be managed and effectively processed and analyzed in order to be useful. Thus, a key component of our program involves the development of the informatic tools necessary to make the data more broadly available and for extracting knowledge and new biological insights from complex data sets.

The development of this new technology is proceeding in concert with its applications to a number of microbial systems (initially *Shewanella oneidensis* MR1, *Deinococcus radiodurans* R1 and *Rhodopseudomonas palustris*) in collaboration with leading experts on each organism. This research is providing the first comprehensive information on the nature of expressed proteins by these systems and how they respond to mutations in the organism or perturbations to its environment. Initial studies applying these approaches have demonstrated the capability for automated high-confidence protein identifications, broad and unbiased proteome coverage, and the capability for exploiting stable-isotope (e.g. ^{15}N) labeling methods to obtain high precision relative protein abundance measurements from microbial cultures. These initial efforts have demonstrated the most complete protein coverage yet obtained for a number of microorganisms, and have begun revealing new biological understandings.

Finally, it is projected that the AMT tag approach will be applicable for making much faster and comprehensive 'metabolome' measurements, and can also likely be extended to the characterization of proteomes (and metabolomes) of much more complex microbial communities.

This research is 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.

B61

New Developments in Statistically Based Methods for Peptide Identification via Tandem Mass Spectrometry

Kenneth D. Jarman, Kristin H. Jarman, Alejandro Heredia-Langner, and **William R. Cannon** (William.Cannon@pnl.gov)

Pacific Northwest National Laboratory, Richland, WA 99352

High-throughput proteomic technologies seek to characterize the state of the proteome in a cell population in much the same manner that DNA microarrays seek to characterize the state of gene expression in a cell population. Characterization of the proteins can be done using several different methods, one of which is to digest the proteins first, typically using trypsin, into peptides which are then analyzed using tandem mass spectrometry (MS/MS). A typical procedure may involve extracting cellular proteins followed by tryptic digestion and then separating the peptides with liquid chromatography. The separated peptides are then identified by MS/MS. Ideally, peptides will subsequently be quantitated, post-translational modifications will be determined and the information regarding the peptides will be assembled into a picture of the proteomic state of a cell population.

Just as with DNA microarrays, quality assurance of the high-throughput process is of paramount importance in order for proteomics to be of value to biologists. If peptides are initially identified poorly, then this information and the information on post-translational state and quantitation of protein expression is not of much value. For this reason, there has been much work recently on developing peptide identification methods for MS/MS spectra. This area of research has proceeded on two fronts, the first of which seeks to take advantage of the wide availability of genome sequences. The database search methods try to identify the peptide that resulted in the observed MS/MS spectrum by picking the best candidate from a list of peptides generated from the genome sequence. De novo methods on the other hand, seek to sequence and hence identify a peptide simply from the observed MS/MS spectrum. Regardless of which approach is used, it is essential to have a method for scoring each peptide so that accurate and reliable identifications can be made.

In this work, we present a statistically rigorous scoring algorithm for peptide identification that can be used alone, or incorporated into a database search algorithm or a de novo peptide sequencing algorithm. Our approach is based on a probabilistic model for the occurrence of spectral peaks corresponding to key partial peptide ion types. In particular, the ion frequencies for the most frequently observed ion types are initially estimated from a training dataset of known sequences. These frequencies are then used to construct a fingerprint for any candidate peptide of interest, where the fingerprint consists of a list of spectral peaks and their corresponding probabilities of appearance. A spectrum is then scored against the candidate fingerprints using a likelihood ratio between the hypothesis that the candidate peptide is not present and the hypothesis that the candidate peptide is present. This likelihood ratio can be used for peptide identification. In addition, a probabilistic score that estimates the probability of a candidate peptide being present in the test sample can be constructed from the likelihood ratio. This approach is tested using a large dataset of over 2000 spectra for tryptic peptides of different lengths ranging from 6-mer to 30-mer amino acids. Performance results indicate that this approach is accurate, and consistent across different peptide lengths and experimental conditions. False positive and false negative error rates for sequence length 10-mer and shorter are generally below 5%, while error rates for sequences longer than 10-mers are typically below 3%.

In addition, we present a Genetic Algorithm (GA) for de novo peptide sequencing. Unlike other de novo construction techniques, this methodology does not try to build amino acid chains by piecing together a feasible path through a graph using the spectral information available but starts with complete sequences and attempts to gradually find one that matches the target spectrum optimally. Due to its building approach, the GA is not immediately deterred by incomplete spectra, peaks produced by unusually occurring peptide fragments or background noise.

Appendix 1: Attendees List

Attendees list as of January 31, 2003.

Eivind Almaas
University of Notre Dame
225 Nieuwland Science Hall
Notre Dame, IN 46556
574-631-3368, Fax 574-631-5952
Almaas.1@nd.edu

Gary Andersen
Lawrence Berkeley National Laboratory
1 Cyclotron Road, Mail Stop 70A-3317
Berkeley, CA 94720
510-495-2795, Fax 510-486-7152
GLAndersen@lbl.gov

Carl Anderson
Brookhaven National Laboratory
Biology Department
Upton, NY 11973
631-344-3375, Fax 631-344-6398
cwa@bnl.gov

Gordon Anderson
Pacific Northwest National Laboratory
P.O. Box 999 MS K8-91
Richland, WA 99352
509-376-9558
gordon@pnl.gov

Adam Arkin
Lawrence Berkeley National Laboratory
1 Cyclotron Road MS 3-144
Berkeley, CA 94720
510-495-2366, Fax 510-486-6059
aparkin@lbl.gov

Daniel Arp
Oregon State University
2082 Cordley
Corvallis, OR 97331
541-737-1294, Fax 541-737-5310
arpd@bcc.orst.edu

Nitin Baliga
Institute for Systems Biology
1441 North 34th Street
Seattle, WA 98118
206-732-1266, Fax 206-732-1299
nbaliga@systemsbiology.org

Jill Banfield
University of California Berkeley
McCone Hall
Berkeley, CA 94720
510-642-9488
jill@seismo.berkeley.edu

Albert-László Barabás
University of Notre Dame
Department of Physics
Notre Dame, IN 46556
574-631-5767, Fax 574-631-5952
alb@nd.edu

John Battista
Louisiana State University and A & M College
Department of Biological Sciences
Baton Rouge, LA 70803
225-578-2810, Fax 225-578-2597
jbattis@lsu.edu

Paul Bayer
U.S. Department of Energy
Biological and Environmental Research
SC-75, GTN,
1000 Independence Ave., SW
Washington, DC 20585-1290
301-903-5324, Fax 301-903-8519
paul.bayer@science.doe.gov

Alex Beliaev
Pacific Northwest National Laboratory
509-376-4183
alexander.beliaev@pnl.gov

Appendix 1: Attendees List

Jeffrey Blanchard
National Center for Genome Resources
2935 Rodeo Park Dr.
Santa Fe, NM 87505
505-995-4405, Fax 505-995-4432
jlb@ncgr.org

Harvey Bolton, Jr
Battelle/Pacific Northwest National Laboratory
P.O. Box 999, Mailstop P7-50
Richland, WA 99352
509-376-3950, Fax 509-376-1321
harvey.bolton@pnl.gov

Daniel Bond
University of Massachusetts at Amherst
413-545-1048
dbond@nre.umass.edu

Mark Borodovsky
Georgia Tech
Atlanta, GA 30332-0230
404-894-8432
mark@amber.gatech.edu

Randy Brich
U.S. Department of Energy
Pacific Northwest National Laboratory
Site Office P.O. Box 550
Richland, WA 99352
509-372-4617, Fax 509-372-4037
randall_f_brich@rl.gov

Fred Brockman
Pacific Northwest National Laboratory
509-376-1252
fred.brockman@pnl.gov

Cindy Bruckner-Lea
Pacific Northwest National Laboratory
P.O. Box 999, Mailstop K8-93
Richland, WA 99352
509-376-2175, Fax 509-376-1044
cindy.bruckner-lea@pnl.gov

Michelle Buchanan
Oak Ridge National Laboratory
Oak Ridge, TN 37830
865-574-4986
buchananmv@ornl.gov

William Cannon
Pacific Northwest National Laboratory
902 Battelle Blvd.
Richland, WA 99352
509-375-6732, Fax 509-375-6631
william.cannon@pnl.gov

David Case
The Scripps Research Institute
Dept. of Molecular Biology
TPC15 La Jolla, CA 92037
858-784-9768, Fax 858-784-8896
case@scripps.edu

Denise Casey
Oak Ridge National Laboratory
Human Genome Management
Information System
1060 Commerce Park, MS 6480
Oak Ridge, TN 37830
865-574-0597, Fax 865-574-9888
caseydk@ornl.gov

Swapnil Chhabra
Sandia National Labs
MO51, Rm122
Biosystems Research Department
Oakland, CA 94610
925-294-4551
swapnil_chhabra@yahoo.com

Sallie Chisholm
Massachusetts Institute of Technology
48-425 MIT
Cambridge, MA 02139
617-253-1771, Fax 617-258-7009
chisholm@mit.edu

Parag Chitnis
National Science Foundation
4201 Wilson Blvd
Arlington, VA 22203
703-292-8443, Fax 703-292-9061
pchitnis@nsf.gov

Linda Chrisey
Office of Naval Research
800 N. Quincy Street
Arlington, VA 22217-5660
703-696-4504, Fax 703-696-1212
chrisel@onr.navy.mil

George Church
Harvard University
200 Longwood Ave.
Boston, MA 02115
617-432-1278, Fax 617-432-7266
church@arep.med.harvard.edu

Stacy Ciufo
University of Massachusetts
Microbiology/ Morrill
4 North Amherst, MA 01003
413-577-1392, Fax 413-545-1578
sciufo@microbio.umass.edu

Dean Cole
U.S. Department of Energy
Medical Science Div., SC-73
Washington, DC 20505
301-903-3268, Fax 301-903-0567
dean.cole@science.doe.gov

James Cole
Michigan State University
Ribosomal Database Project-II,
2225A Biomedical Physical Sciences
East Lansing, MI 48824
517-353-3843, Fax 517-353-2917
colej@msu.edu

Steve Colson
Pacific Northwest National Laboratory
P.O. Box 999, MS-IN K8-88
Richland, WA 99352
509-376-4598, Fax 509-376-0846
steven.colson@pnl.gov

Michael Colvin
Lawrence Livermore National Laboratory
Mailstop L-448
Livermore, CA 94552
925-423-9177, Fax 925-424-6605
colvin2@LLNL.gov

Maddalena Coppi
University of Massachusetts at Amherst
Department of Microbiology
203N Morrill IVN
Amherst, MA 01003
413-545-2067
mcoppi@microbio.umass.edu

Bob Coyne
National Science Foundation
703-292-8439, Fax 703-292-9061
rcoyne@nsf.gov

Terence Critchlow
Lawrence Livermore National Laboratory
7000 East Ave MS L-560
Livermore, CA 94550
925-423-5682
critchlow@llnl.gov

Claribel Cruz-Garcia
Michigan State University
Center for Microbial Ecology Plant
and Soil Science
Building Rm 540
East Lansing, MI 48824
517-353-7858, Fax 517-353-2917
cruzgarc@msu.edu

Brian Davison
Oak Ridge National Laboratory
P.O. Box 2009
Oak Ridge, TN 37831-6226
865-576-8522, Fax 865-574-6442
davisonbh@ornl.gov

Jonathan Delatizky
BBN Technologies
10 Moulton Street
Cambridge, MA 02138
617-873-3366
delatizky@bbn.com

Ed DeLong
MBARI
7700 Sandholdt
Moss Landing, CA 95039
831-775-1843, Fax 831-775-1646
delong@mbari.org

David Dixon
Pacific Northwest National Laboratory
P.O. Box 999, MS K9-90
Richland, WA 99352
509-372-4999, Fax 509-375-6776
david.dixon@pnl.gov

Appendix 1: Attendees List

Mitchel Doktycz
Oak Ridge National Laboratory
P.O. Box 2008, MS 6123
Oak Ridge, TN 37831-6123
865-574-6204, Fax 865-574-6210
doktyczmj@ornl.gov

Timothy Donohue
University of Wisconsin-Madison
1550 Linden Drive
Madison, WI 53562
608-262-4663, Fax 608-262-9865
tdonohue@bact.wisc.edu

Janet Dorigan
U.S. Government
703-918-9672
jdorigan@aol.com

Norman Dovichi
University of Washington
Department of Chemistry
Seattle, WA 98195-1700
206-543-7835, Fax 206-685-8665
dovichi@chem.washington.edu

Kenneth Downing
Lawrence Berkeley National Laboratory
Donner Lab
Berkeley, CA 94720
510-486-5941, Fax 510-486-6488
khdowning@lbl.gov

Daniel Drell
U.S. Department of Energy
SC-72/GTN 1000 Independence Ave., SW
Washington, DC 20585-1290
301-903-4742, Fax 301-903-8521
daniel.drell@science.doe.gov

Leland Ellis
USDA/ARS
National Program Staff
5601 Sunnyside Avenue
Beltsville, MD 20705-5138
301-504-4788, Fax 301-504-4725
lce@ars.usda.gov

Brendlyn Faison
U.S. Department of Energy
Office of Biological and Environmental Research
301-903-0042, Fax 301-903-4154
brendlyn.faison@science.doe.gov

Jordan Feidler
The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102-7508
703-883-5624, Fax 703-883-6501
feidler@mitre.org

Brad Fenwick
USDA Competitive Programs
785-532-4412
fenwick@vet.ksu.edu

Matthew Fields
Oak Ridge National Laboratory
1 Bethel Valley Road
Oak Ridge, TN 37831-6038
865-241-3775, Fax 865-576-8646
fieldsm1@ornl.gov

Marvin Frazier
U.S. Department of Energy
Office of Biological and Environmental Research
19901 Germantown Rd.
Germantown, MD 20874
301-903-5468, Fax 301-903-8521
marvin.frazier@science.doe.gov

Jim Fredrickson
Pacific Northwest National Laboratory
P.O. Box 999, Mailstop P7-50
Richland, WA 99352
509-376-7063, Fax 509-376-9650
jim.fredrickson@pnl.gov

George Garrity
Michigan State University
6162 Biomedical Physical Sciences Bldg.
East Lansing, MI 48824-4320
517-432-2459, Fax 517-432-2458
garrity@msu.edu

Al Geist
Oak Ridge National Laboratory
P.O. Box 2008
Oak Ridge, TN 37831-6367
865-574-3153, Fax 865-574-0680
gst@ornl.gov

Svetlana Gerdes
Integrated Genomics, Inc.
2201 W. Campbell Park Drive
Chicago, IL 60612
312-491-0846, Fax 312-491-0856
sveta@integratedgenomics.com

Raymond Gesteland
University of Utah
Department of Human Genetics
15 N. 2030 E., Rm. 7410
Salt Lake City, UT 84112-5330
801-581-5190, Fax 801-585-3910
ray.gesteland@genetics.utah.edu

Steven Gill
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-315-2521
srgill@tigr.org

Carol Giometti
Argonne National Laboratory
9700 South Cass Avenue,
Building 202
Argonne, IL 60439
630-252-3839, Fax 630-252-3387
csgiometti@anl.gov

Silvia Gonzalez-Acinas
Massachusetts Institute of Technology
15 Vassar St., 48-108
Cambridge, MA 02139
617-253-7651
sacinas@mit.edu

Andrey Gorin
Oak Ridge National Laboratory
1 Bethel Valley Road
Bld. 6012, MS 6367
Oak Ridge, TN 37831
865-241-3972, Fax 865-574-0680
agor@ornl.gov

Debbie Gracio
Pacific Northwest National Laboratory
P.O. Box 999 MS K1-85
Richland, WA 99352
509-375-6362, Fax 509-375-6631
gracio@pnl.gov

Yonatan Grad
Church Lab, Harvard Medical School
250 Longwood Avenue, Rm 221
Seeley Mudd Building
Boston, MA 02115
617-432-0063, Fax 617-432-0065
yonatan_grad@student.hms.harvard.edu

Masood Hadi
Sandia National Laboratories
P.O. Box 969, MS 9951
7011 East Ave
Livermore, CA 94550
925-294-4893
mzhadi@sandia.gov

Bruce Hamilton
National Science Foundation
4201 Wilson Boulevard, Suite 565
Arlington, VA 22230
703-292-7066
bhamilto@nsf.gov

Frank Harris
Oak Ridge National Laboratory
P.O. Box 2008
Oak Ridge, TN 37831
865-574-4333, Fax 865-574-9869
harrisf@ornl.gov

Caroline Harwood
University of Iowa
Department of Microbiology
3-432 BSB
Iowa City, IA 52242
319-335-7783, Fax 319-335-7679
caroline-harwood@uiowa.edu

Terry Hazen
Lawrence Berkeley National Laboratory
MS 70A-3317
One Cyclotron Rd.
Berkeley, CA 94720
510-486-6223, Fax 510-486-7152
tchazen@lbl.gov

Grant Heffelfinger
Sandia National Laboratories
P.O. Box 5800, MS-0885
Albuquerque, NM 87185
505-845-7801, Fax 505-284-3093
gsheffe@sandia.gov

Appendix 1: Attendees List

Diane Hevehan
U.S. Department of Defense
3050 Defense Pentagon, Rm. 3C257
Washington, DC 20301
703-693-6835, Fax 703-695-0476
diane.hevehan@osd.mil

Ed Hildebrand
White House Office of Science
and Technology Policy
1801 Pennsylvania Ave. NW
Washington, DC 20502
202-456-7341, Fax 202-456-6027
childebr@ostp.eop.gov

Roland F. Hirsch
U.S. Department of Energy
Medical Sciences Division,
SC-73 GTN 1000 Independence Avenue SW
Washington, DC 20585-1290
301-903-9009
roland.hirsch@science.doe.gov

Lynette Hirschman
MITRE
202 Burlington Rd
Bedford, MA 01730
781-271-7789, Fax 781-271-2780
lynette@mitre.org

Hoi-Ying Holman
Lawrence Berkeley National Laboratory
One Cyclotron Road
Berkeley, CA 94720
510-486-5943, Fax 510-486-7152
hyholman@lbl.gov

John Houghton
U.S. Department of Energy
Biological and Environmental Research
1000 Independence Avenue, SW
Washington, DC 20585-1290
301-903-8288, Fax 301-903-8521
john.houghton@science.doe.gov

Peter Hoyt
Oak Ridge National Laboratory
P.O. Box 2008, MS 6123
Oak Ridge, TN 37831-6123
865-574-6211, Fax 865-574-6210
hoytpr@ornl.gov

Greg Hurst
Oak Ridge National Laboratory
P.O. Box 2008, MS 6131
Oak Ridge, TN 37831
865-574-6142, Fax 865-576-8559
hurstgb@ornl.gov

Barbara Jasny
SCIENCE/AAAS
202-326-6515
bjasny@aaas.org

Gary Johnson
U.S. Department of Energy
Office of Science
301-903-5800
Gary.Johnson@science.doe.gov

Matthew Kane
National Science Foundation
4201 Wilson Blvd.
Arlington, VA 22306
703-292-7186
mkane@nsf.gov

Arthur Katz
U.S. Department of Energy
Office of Biological and Environmental Research
1000 Independence Avenue, SW
Washington, DC 20585-11290
301-903-4932, Fax 301-903-8521
arthur.katz@science.doe.gov

Jay Keasling
University of California
201 Gilman Hall
Dept. of Chemical Engineering
Berkeley, CA 94720-1462
510-642-4862, Fax 510-642-7657
keasling@socrates.berkeley.edu

Martin Keller
Diversa Corporation
4955 Directors Place
San Diego, CA 92121
858-526-5162
mkeller@diversa.com

Steve Kennel
Oak Ridge National Laboratory
P.O. Box 2008, MS 6101
Oak Ridge, TN 37831-6101
865-574-0825, Fax 865-574-6210
kennelsj@ornl.gov

David Kirchman
University of Delaware
700 Pilottown Road
Lewes, DE 19958
302-645-4375, Fax 302-645-4028
kirchman@udel.edu

Joel Klappenbach
Center for Microbial Ecology
Michigan State University
540 Plant & Soil Sciences
East Lansing, MI 48824
517-353-9021, Fax 517-353-2917
klappenb@msu.edu

Michael Knotek
Consultant
10127 N. Bighorn Butte Drive
Oro Valley, AZ 85737
520-591-8108, Fax 520-877-3233
m.knotek@verizon.net

Eugene Kolker
BIATECH
19310 North Creek Parkway, Ste. 115
Bothell, WA 98011
425-481-7200 ext 100, Fax 425-481-5384
ekolker@bitech.org

Julia Krushkal
University of Tennessee Health Science Center
Department of Preventive Medicine
66 N. Pauline, Ste. 615
Memphis, TN 38163
901-448-1361, Fax 901-448-7041
jkrushka@utmem.edu

Henrietta Kulaga
Contractor-IPTO
20 Firstfield Rd. Suite 125
Gaithersburg, MD 20878
301-990-9570, Fax 301-990-3680
hkulaga@snap.org

Sri Kumar
DARPA
3701 N. Fairfax
Arlington, VA 22203
571-218-4275
skumar@darpa.mil

Cheryl Kuske
Los Alamos National Laboratory
Bioscience Division, M888
Los Alamos, NM 87545
505-665-4800, Fax 505-665-3024
kuske@lanl.gov

Todd Lane
Sandia National Laboratories
P.O. Box 969 MS9951
Livermore, CA 94550-969
925-294-2057, Fax 925-294-3020
twlane@sandia.gov

Frank Larimer
Oak Ridge National Laboratory
1060 Commerce Park Drive
Oak Ridge, TN 37831
865-574-1253, Fax 865-576-5332
larimerfw@ornl.gov

Michael Laub
Harvard University
Bauer Center for Genomics Research
7 Divinity Ave
Cambridge, MA 02138
617-384-9647
laub@cgr.harvard.edu

Charles Lawrence
Wadsworth Center
Empire State Plaza
Albany, NY 12201
518-473-3382, Fax 518-474-7992
lawrence@wadsworth.org

John Leigh
University of Washington
Campus Box 357242
Seattle, WA 98195-7242
206-685-1390, Fax 206-543-8297
leighj@u.washington.edu

Appendix 1: Attendees List

Jerry Li
National Institute of General Medical Sciences
45 Center Dr., Rm 2As.19F
Bethesda, MD 20892
301-594-0682, Fax 301-480-2004
lij@nigms.nih.gov

Tim Lilburn
American Type Culture Collection
10801 University Boulevard
Manassas, VA 20110
703-365-2700 x599, Fax 703-365-2740
tlilburn@atcc.org

Debbie Lindell
Massachusetts Institute of Technology
MIT 48-336, 77 Massachusetts Ave
Cambridge, MA 02139
617-258-6835, Fax 617-258-7009
dlindell@mit.edu

Mary Lipton
Pacific Northwest National Laboratory
3335 Q. Avenue
Richland, WA 99352
509-373-9039, Fax 509-376-7722
mary.lipton@pnl.gov

Derek Lovley
Department of Microbiology
University of Massachusetts
Amherst, MA 01003
413-545-9651, Fax 413-545-1578
dlovley@microbio.umass.edu

Susan Lucas
DOE Joint Genome Institute
2800 Mitchell Drive
Walnut Creek, CA 94598
925-296-5638, Fax 925-296-5875
lucas11@llnl.gov

Diane Makowski
Argonne National Laboratory
9700 S. Cass Avenue
Argonne, IL 60439
630-252-3963, Fax 630-252-5517
drodi@anl.gov

Stephanie Malfatti
Lawrence Livermore National Laboratory
P.O. Box 808
Livermore, CA 94550
925-424-6274, Fax 925-422-2099
malfatti3@llnl.gov

Julie Malicoat
Oak Ridge Associated Universities
865-576-9952
malicoaj@orau.gov

Reinhold Mann
Pacific Northwest National Laboratory
P.O. Box 999
Richland, WA 99352
509-376-6764, Fax 509-375-6844
mannrc@pnl.gov

Betty Mansfield
Oak Ridge National Laboratory
1060 Commerce Park
Oak Ridge, TN 37830
865-576-6669, Fax 865-574-9888
mansfieldbk@ornl.gov

Terence Marsh
Michigan State University
41 Giltner Hall
East Lansing, MI 48824
517-432-1365, Fax 517-432-3770
MARSHT@msu.edu

Vincent Martin
Lawrence Berkeley National Laboratory
201 Gilman Hall
Berkeley, CA 94708
510-642-9506
vincentm@socrates.berkeley.edu

Anthony Martino
Sandia National Laboratories
P.O. Box 969, MS9951
Livermore, CA 94551
925-294-2095, Fax 925-294-3020
martino@sandia.gov

Lee Ann McCue
Wadsworth Center
P.O. Box 509
Albany, NY 12201
518-473-3382, Fax 518-473-2900
mccue@wadsworth.org

Shawn McLaughlin
National Oceanic and Atmospheric Administration
904 S. Morris St.
Oxford, MD 21654
410-226-5193, Fax 410-226-5925
shawn.mclaughlin@noaa.gov

Barbara Methé
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-0200, Fax 301-838-0208
bmethe@tigr.org

Noelle Metting
U.S. Department of Energy
SC-72/GTN 1000 Independence Ave., SW
Washington, DC 20585-1290
301-903-8309
noelle.metting@science.doe.gov

Simon Minovitsky
Lawrence Berkeley Laboratory
510-495-2913
sminovitsky@lbl.gov

Bud Mishra
Courant Institute and
Cold Spring Harbor Laboratory
212-998-3464
mishra@nyu.edu

Julie Mitchell
San Diego Supercomputer Center
University of California San Diego
9500 Gilman Dr. MC0527
La Jolla, CA 92103
858-534-5126, Fax 858-822-3631
mitchell@sdsc.edu

Emmanuel Mongodin
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-5859, Fax 301-838-0208
emongodin@tigr.org

Sue Morss
Argonne National Laboratory
9700 South Cass Avenue
Argonne, IL 60439
630-252-6784, Fax 630-252-6720
smorss@anl.gov

Ali Navid
Indiana University
Chemistry Building
800 E. Kirkwood Ave
Bloomington, IN 47405
812-855-2047, Fax 812-855-8300
anavid@indiana.edu

Karen Nelson
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-3565, Fax 301-838-0208
kenelson@tigr.org

Camilla Nesbo
Dalhousie University
5850 College Street
Halifax, Nova Scotia,
Canada B3H 1X5
902-494-2968, Fax 902-494-1355
cnesbo@dal.ca

Frank Olken
Lawrence Berkeley National Laboratory
1 Cyclotron Rd.
Mailstop 50B-3238
Berkeley, CA 94720-8147
510-486-5891, Fax 510-486-4004
olken@lbl.gov

Peter Ortoleva
Indiana University
Chemistry Building
800 E. Kirkwood
Bloomington, IN 47405
812-855-2717, Fax 812-855-8300
ortoleva@indiana.edu

Ross Overbeek
Integrated Genomics, Inc.
2201 W. Campbell Park Dr.
Chicago, IL 60612
630-567-7677, Fax 312-491-0856
Ross@IntegratedGenomics.com

Appendix 1: Attendees List

Himadri Pakrasi
Washington University
Department of Biology
Box 1137
St. Louis, MO 63130
314-935-6853, Fax 314-935-6803
pakrasi@biology2.wustl.edu

Anna Palmisano
U.S. Department of Energy
19901 Germantown Rd.
Germantown, MD 20852
301-903-9963, Fax 301-903-8519
anna.palmisano@science.doe.gov

Morey Parang
ApoCom Genomics
11020 Solway School Rd., Suite 101
Knoxville, TN 37931
865-927-6120, Fax 865-927-6122
mparang@apocom.com

Bahram Parvin
Lawrence Berkeley National Laboratory
M.S. 50B-2239
Berkeley, CA 94720
510-486-6203
parvin@media.lbl.gov

Ari Patrinos
U.S. Department of Energy
Office of Biological and Environmental Research
SC-70/GTN 1000 Independence Ave., SW
Washington, DC 20585-1290
301-903-3251, Fax 301-903-5051
ari.patrinos@science.doe.gov

Deborah Payne
Pacific Northwest National Laboratory
902 Battelle Boulevard
Richland, WA 99352
509-375-2904, Fax 509-375-6631
debbie.payne@pnl.gov

Dale Pelletier
Oak Ridge National Laboratory
P.O. Box 2008, MS 6149
Oak Ridge, TN 37831
865-574-1239
pelletierda@ornl.gov

Martin Polz
Massachusetts Institute of Technology
77 Mass. Ave, 48-421
Cambridge, MA 02140
617-253-7128
mpolz@mit.edu

Valerie Reed
U.S. Department of Energy
Office of Biomass Programs
1000 Independence Ave, SW
Washington, DC 20585
202-586-5618
valerie.sarisky-reed@hq.doe.gov

Haluk Resat
Pacific Northwest National Laboratory
P.O. Box 999, Mail Stop K1-83
Richland, WA 99352
509-372-6340, Fax 509-375-6631
haluk.resat@pnl.gov

Paul Richardson
DOE Joint Genome Institute
2800 Mitchell Drive
Walnut Creek, CA 94598
925-296-5851, Fax 925-296-5875
pmricharsond

Mark Rintoul
Sandia National Laboratories
505-844-9592
rintoul@sandia.gov

Robert Roberson
Arizona State University
Department of Plant Biology
Tempe, AZ 85287-1601
480-965-8618
Robert.Roberson@asu.edu

Karin Rodland
Pacific Northwest National Laboratory
P.O. Box 999
Richland, WA 99352
509-376-7605, Fax 509-376-6767
karin.rodland@pnl.gov

Lars Rohlin
University of California Los Angeles
405 Hilgard Ave, BH5531
Los Angeles, CA 90095
310-825-5849, Fax 310-206-4107
lrohlin@ucla.edu

Charles Romine
ASCR/MICS
U.S. Department of Energy
1000 Independence Ave., SW
Washington, DC 20585-1290
301-903-5152, Fax 301-903-7774
romine@er.doe.gov

Margie Romine
Pacific Northwest National Laboratory
P.O. Box 999, MS P7-50
Richland, WA 99352
509-376-8287, Fax 509-376-9650
margie.romine@pnl.gov

Lucian Russell
Expert Reasoning & Decisions LLC
6012 Jewell Court
Alexandria, VA 22312
703-916-0474, Fax 703-642-0954
lucianr@verizon.net

Andrey Rzhetsky
Columbia University
1150 St. Nicholas Ave.
Russ Berrie Pavilion
New York, NY 10032
212-851-5150, Fax 212-851-5149
ar345@columbia.edu

Herbert Sauro
Keck Graduate Institute
535 Watson Drive
Claremont, CA 91711
909-607-0377, Fax 909-607-8598
hsauro@kgi.edu

Christophe Schilling
Genomatica, Inc.
5405 Morehouse Drive
San Diego, CA 92121
858-362-8550, Fax 858-824-1772
cschilling@genomatica.com

David Schwartz
University of Wisconsin-Madison
Genetics/Biotechnology Center
425 Henry Mall
Madison, WI 53706
608-265-0546
dcschwartz@facstaff.wisc.edu

Margrethe Serres
Marine Biological Laboratory
7 MBL Street
Woods Hole, MA 02543
508-289-7388, Fax 508-457-4727
mserres@mbl.edu

Imran Shah
University of Colorado
4200 E 9th Ave, B-119
Denver, CO 80262
303-315-7222, Fax 303-315-3183
imran.shah@uchsc.edu

Henry Shaw
U.S. Department of Energy
Office of Biological and Environmental Research
ERSD
301-903-3947
henry.shaw@science.doe.gov

Arie Shoshani
Lawrence Berkeley National Laboratory
1 Cyclotron Rd.
Berkeley, CA 94720
510-486-5171, Fax 510-486-4004
shoshani@lbl.gov

Robert Siegel
Pacific Northwest National Laboratory
902 Battelle Blvd
Richland, WA 99352
509-372-6765
robert.siegel@pnl.gov

Richard Smith
Pacific Northwest National Laboratory
P.O. Box 999 (K8-98)
Richland, WA 99352
509-376-0723, Fax 509-376-7722
rds_smith@pnl.gov

Appendix 1: Attendees List

David Springer
Pacific Northwest National Laboratory
Biological Sciences Division, K4-12
Richland, WA 99352
509-372-6762, Fax 509-372-6544
David.Springer@pnl.gov

Thomas Squier
Pacific Northwest National Laboratory
P.O. Box 999, MS P7-53
Richland, WA 99352
509-376-2218, Fax 509-376-1321
Thomas.Squier@pnl.gov

David A. Stahl
University of Washington
Dept Civil & Environ. Eng.
302 More Hall, Box 352700
Seattle, WA 98195-2700
206-685-3464, Fax 206-685-9185
dastahl@u.washington.edu

Marvin Stodolsky
U.S. Department of Energy
Office of Biological and Environmental Research
1000 Independence Avenue, SW
SC-72/GTN
Washington, DC 20585-1290
301-903-4475, Fax 301-903-8521
Marvin.Stodolsky@science.doe.gov

T.P. Straatsma
Pacific Northwest National Laboratory
P.O. Box 999, MS K1-83
Richland, WA 99352
509-375-2802, Fax 509-375-6631
tps@pnl.gov

Ray Stults
Los Alamos National Laboratory
505-667-5535, Fax 505-667-1113
rstults@lanl.gov

F. Robert Tabita
Department of Microbiology
Ohio State University
484 West 12th Avenue
Columbus, OH 43210-1292
614-292-4297, Fax 614-292-6337
tabita.1@osu.edu

David Thomassen
U.S. Department of Energy
Office of Biological and Environmental Research
1000 Independence Avenue, SC-72
Germantown Building
Washington, DC 20585-1290
301-903-9817, Fax 301-903-8521
david.thomassen@science.doe.gov

Dorothea Thompson
Oak Ridge National Laboratory
1 Bethel Valley Road
Oak Ridge, TN 37831-6038
865-574-4815, Fax 865-576-8646
thompsondk@ornl.gov

Jerilyn Timlin
Sandia National Laboratories
P.O. Box 5800, M.S. 0886
Albuquerque, NM 87111
505-844-7932
jatimli@sandia.gov

Cary Tuckfield
Savannah River Technology Center
Bldg. 773-42A
Aiken, SC 29808
803-725-8215, Fax 803-725-8829
cary.tuckfield@srs.gov

Wim Vermaas
Arizona State University
Department of Plant Biology
Box 871601
Tempe, AZ 85287-1601
480-965-3698, Fax 480-965-6899
wim@asu.edu

Michael Viola
U.S. Department of Energy
SC-70/GTN 1000 Independence Ave., SW
Washington, DC 20585-1290
301-903-5346, Fax 301-903-0567
michel.viola@science.doe.gov

Lawrence Wackett
University of Minnesota
Biochemistry
1479 Gortner Avenue
St. Paul, MN 55108
612-625-3785, Fax 612-625-1700
wackett@biosci.cbs.umn.edu

Lisa Waidner
University of Delaware
College of Marine Studies
700 Pilottown Rd.
Lewes, DE 19958
302-645-4030, Fax 302-645-4028
lwaidner@udel.edu

Elizabeth Weitzke
Indiana University
Chemistry Building
800 E. Kirkwood Ave.
Bloomington, IN 47405
812-855-2047, Fax 812-855-8300
eweitzke@indiana.edu

Owen White
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-5824, Fax 301-838-0208
owhite@tigr.org please cc jfowler@tigr.org

Julian Whitelegge
University of California Los Angeles
Dept of Chemistry
405 Hilgard Ave
Los Angeles, CA 90095
310-794-5156, Fax 310-206-2161
jpw@chem.ucla.edu

William Whitman
University of Georgia
Dept. of Microbiology
Athens, GA 30602-2605
706-542-4219, Fax 706-542-2674
whitman@arches.uga.edu

Steven Wiley
Pacific Northwest National Laboratory
902 Battelle Boulevard
Richland, WA 99352
509-373-6218, Fax 509-376-1494
steven.wiley@pnl.gov

David Wilson
Cornell University
458 Biotechnology Bldg.
Ithaca, NY 14853
607-255-5706, Fax 607-255-2428
dbw3@cornell.edu

Ying Xu
Oak Ridge National Laboratory
1060 Commerce Park Drive, MS 6480
Life Sciences Division
Oak Ridge, TN 37830-6480
865-574-7263, Fax 865-241-1965
xyn@ornl.gov

Huei-Che Bill Yen
University of Missouri-Columbia
Rm 117, Schweitzer Hall
Columbia, MO 65211
573-882-9771, Fax 573-882-5635
yenh@missouri.edu

Jizhong Zhou
Oak Ridge National Laboratory
1 Bethel Valley Road
Oak Ridge, TN 37831-6038
865-576-7544, Fax 865-576-8646
zhouj@ornl.gov

Shiguo Zhou
University of Wisconsin-Madison
425 Henry Mall
Madison, WI 53706
608-265-7930
szhou@lmcg.wisc.edu

Robin Zimmer
ApoCom Genomics
11020 Solway School Rd.
Suite 101
Knoxville, TN 37931
865-927-6120, Fax 865-927-6122
robzimmer@apocom.com

Appendix 2: Poster Presenters

Gary Andersen
- **B51** Development of Vectors for Detecting Protein-Protein Interactions in Bacteria

Adam Arkin
- **A4** Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria

Dan Arp
- **B13** Gene Expression Profiles in *Nitrosomonas europaea*, an Obligate Chemolithoautotroph

Jillian F. Banfield
- **B7** Ecological and Evolutionary Analyses of a Spatially and Geochemically Confined Acid Mine Drainage Ecosystem Enabled by Community Genomics

Albert-László Barabási
- **A26** Hierarchical Organization of Modularity in Metabolic Networks

John R. Battista
- **B9** Uncovering the Regulatory Networks Associated with Ionizing Radiation-Induced Gene Expression in *D. radiodurans* R1

Mark Borodovsky
- **A54** Predicting Genes from Prokaryotic Genomes: Are "Atypical" Genes Derived from Lateral Gene Transfer?

Fred Brockman
- **B5** Approaches for Obtaining Genome Sequence from Contaminated Sediments Beneath a Leaking High-Level Radioactive Waste Tank

Michelle Buchanan
- **A12** New Approaches for High-Throughput Identification and Characterization of Protein Complexes

William R. Cannon
- **B61** New Developments in Statistically Based Methods for Peptide Identification via Tandem Mass Spectrometry

David A. Case
- **A32** Parallel Scaling in Amber Molecular Dynamics Simulations

Denise Casey
- **B63** Communicating Genomes to Life

George Church
- **A2** Microbial Ecology, Proteogenomics, and Computational Optima

James R. Cole
- **A44** A Web-Based Laboratory Information Management System (LIMS) for Laboratory Microplate Data Generated by High-Throughput Genomic Applications

Michael Colvin
- **A56** Advanced Molecular Simulations of *E. coli* Polymerase III

Michael J. Daly
- **B29** The Dynamics of Cellular Stress Responses in *Deinococcus radiodurans*

Jonathan Delatizky
- **A46** BioSketchpad: An Interactive Tool for Modeling Biomolecular and Cellular Networks

D. A. Dixon
- **A6** Bioinformatics and Computing in the Genomes to Life Center for Molecular and Cellular Systems

Timothy Donohue
- **B41** The Molecular Basis for Metabolic and Energetic Diversity

Norman J. Dovichi
- **B57** Single Cell Proteomics—*D. radiodurans*

Appendix 2: Poster Presenters

Ken Downing
- **B55** Electron Tomography of Whole Bacterial Cells

James K. Fredrickson
- **B21** Environmental Sensing, Metabolic Response and Regulatory Networks in the Respiratory Versatile Bacterium *Shewanella oneidensis* MR-1

George M. Garrity
- **A36** Towards a Self-Organizing and Self-Correcting Prokaryotic Taxonomy

Raymond Gesteland
- **B35** In Search of Diversity: Understanding How Post-Genomic Diversity is Introduced to the Proteome

Carol S. Giometti
- **B37** The Microbial Proteome Project: A Database of Microbial Protein Expression in the Context of Genome Analysis

Carol S. Giometti
- **B39** Analysis of the *Shewanella oneidensis* Proteome in Cells Grown in Continuous Culture

D. K. Gracio
- **A6** Bioinformatics and Computing in the Genomes to Life Center for Molecular and Cellular Systems

Caroline S. Harwood
- **B11** Strategies to Harness the Metabolic Diversity of *Rhodopseudomonas palustris*

Grant S. Heffelfinger
- **A20** Carbon Sequestration in *Synechococcus sp.*: From Molecular Machines to Hierarchical Modeling

P. R. Hoyt
- **A14** Automation of Protein Complex Analyses in *Rhodopseudomonas palustris* and *Shewanella oneidensis*

Gregory B. Hurst
- **A8** Mass Spectrometry in the Genomes to Life Center for Molecular and Cellular Systems

David L. Kirchman
- **B3** A Metagenomic Library of Bacterial DNA Isolated from the Delaware River

Joel A. Klappenbach
- **A44** A Web-Based Laboratory Information Management System (LIMS) for Laboratory Microplate Data Generated by High-Throughput Genomic Applications

Eugene Kolker
- **A64** Interdisciplinary Study of *Shewanella oneidensis* MR-1's Metabolism & Metal Reduction

Cheryl R. Kuske
- **B1** Identification and Isolation of Active, Non-Cultured Bacteria from Radionuclide and Metal Contaminated Environments for Genome Analysis

John Leigh
- **B25** Global Regulation in the Methanogenic Archaeon *Methanococcus maripaludis*

Mary S. Lipton
- **B39** Analysis of the *Shewanella oneidensis* Proteome in Cells Grown in Continuous Culture

Derek Lovley
- **A24** Analysis of the Genetic Potential and Gene Expression of Microbial Communities Involved in the *in situ* Bioremediation of Uranium and Harvesting Electrical Energy from Organic Matter

Derek Lovley
- **A34** Microbial Cell Model of *G. sulfurreducens*: Integration of in silico Models and Functional Genomic Studies

S. Malfatti
- **B33** Finishing and Analysis of the *Nostoc punctiforme* Genome

Betty K. Mansfield
- **B63** Communicating Genomes to Life

Anthony Martino
- **A16** Analysis of Protein Complexes from a Fundamental Understanding of Protein Binding Domains and Protein-Protein Interactions in *Synechococcus* WH8102

Harley McAdams
- **A40** Characterization of Genetic Regulatory Circuitry Controlling Adaptive Metabolic Pathways

Lee Ann McCue
- **A52** Comparative Genomics Approaches to Elucidate Transcription Regulatory Networks

Julie C. Mitchell
- **A48** Molecular Docking with Adaptive Mesh Solutions to the Poisson- Boltzmann Equation

Karen E. Nelson
- **B27** Identification of Regions of Lateral Gene Transfer Across the Thermotogales

Howard Ochman
- **B19** Lateral Gene Transfer and the History of Bacterial Genomes

Peter J. Ortoleva
- **A58** *Karyote*®: Automated Physico-Chemical Cell Model Development Through Information Theory

Brian Palenik
- **A18** Carbon Sequestration in *Synechococcus*: Microarray Approaches

Morey Parang
- **A60** The Commercial Viability of EXCAVATOR™: A Software Tool For Gene Expression Data Clustering

D. A. Payne
- **A6** Bioinformatics and Computing in the Genomes to Life Center for Molecular and Cellular Systems

Haluk Resat
- **A38** Computational Framework for Microbial Cell Simulations

Mark D. Rintoul
- **A22** Systems Biology Models for *Synechococcus sp.*

Diane J. Rodi
- **B49** Identification of ATP Binding Proteins within Sequenced Bacterial Genomes Utilizing Phage Display Technology

Herbert M Sauro
- **A42** Data Exchange and Programmatic Resource Sharing: The Systems Biology Workbench, BioSPICE and the Systems Biology Markup Language (SBML)

Margrethe H. Serres
- **B31** Analysis of Proteins Encoded on the *S. oneidensis* MR-1 Chromosome, Their Metabolic Associations and Paralogous Relationships

Christophe H. Schilling
- **A30** SimPheny: A Computational Infrastructure Bringing Genomes to Life

David C. Schwartz
- **B45** Comparative Optical Mapping: A New Approach for Microbial Comparative Genomics

David C. Schwartz
- **B47** Optical Mapping of Multiple Microbial Genomes

Imran Shah
- **A28** Computational Elucidation of Metabolic Pathways

Richard D. Smith
- **B59** Genomes to Proteomes to Life: Application of New Technologies for Comprehensive, Quantitative and High Throughput Microbial Proteomics

T. P. Straatsma
- **A62** Modeling Electron Transfer in Flavocytochrome c3 Fumarate Reductase

F. Robert Tabita
- **B17** The *Rhodopseudomonas palustris* Microbial Cell Project

Appendix 2: Poster Presenters

Jerilyn Timlin
- **A18** Carbon Sequestration in *Synechococcus*: Microarray Approaches

Wim Vermaas
- **B43** Integrative Studies of Carbon Generation and Utilization in the Cyanobacterium *Synechocystis sp.* PCC 6803

Lawrence P. Wackett
- **A50** Functional Analysis and Discovery of Microbial Genes Transforming Metallic and Organic Pollutants: Database and Experimental Tools

William Whitman
- **B25** Global Regulation in the Methanogenic Archaeon *Methanococcus maripaludis*

H. Steven Wiley
- **A10** Genomes to Life Center for Molecular and Cellular Systems: A Research Program for Identification and Characterization of Protein Complexes

David B. Wilson
- **B15** Genomics of *Thermobifida fusca* Plant Cell Wall Degradating Proteins

Jizhong Zhou
- **B23** Integrated Analysis of Protein Complexes and Regulatory Networks Involved in Anaerobic Energy Metabolism of *Shewanella oneidensis* MR-1

Jizhong Zhou
- **B53** Development and Use of Microarray-Based Integrated Genomic Technologies for Functional Analysis of Environmentally Important Microorganisms

Shiguo Zhou
- **B45** Comparative Optical Mapping: A New Approach for Microbial Comparative Genomics

Shiguo Zhou
- **B47** Optical Mapping of Multiple Microbial Genomes

Robin D. Zimmer
- **A60** The Commercial Viability of EXCAVATOR™: A Software Tool For Gene Expression Data Clustering

Appendix 3: GTL Web Sites

Web Sites

Genomes to Life Web Site

- GTL Roadmap, April 2001: http://DOEGenomesToLife.org
- Image Gallery http://DOEGenomesToLife.org/gallery/images.html
- Payoffs for the Nation http://DOEGenomesToLife.org/payoffs.html
- Current Research http://DOEGenomesToLife.org/research/index.html
- GTL Program FY 2003 Call for Applications http://www.er.doe.gov/production/grants/Fr03-05.html

Current GTL Program Projects

- July 23, 2002, Press Release "Energy Department Awards $103 Million for Post-Genomic Research" http:// DOEGenomesToLife.org/research/2002awards.html
- Oak Ridge National Laboratory for Genomes to Life Center for Molecular and Cellular Systems: A Research Program for Identification and Characterization of Protein Complexes http://www.ornl.gov/GenomestoLife/
- Lawrence Berkeley National Laboratory for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria http://vimss.lbl.gov/
- Sandia National Laboratories for Carbon Sequestration in *Synechococcus*: From Molecular Machines to Hierarchical Modeling http://www.genomes-to-life.org/
- University of Massachusetts, Amherst, for Analysis of the Genetic Potential and Gene Expression of Microbial Communities Involved in the in situ Bioremediation of Uranium and Harvesting Electrical Energy from Organic Matter http://www.geobacter.org
- Harvard Medical School for Microbial Ecology, Proteogenomics and Computational Optima http://arep.med.harvard.edu/DOEGTL/

Complementary Web Sites

- Human Genome Project Information http://www.ornl.gov/hgmis/
- DOE Microbial Genome Program http://www.ornl.gov/microbialgenomes/
- Microbial Genomics Gateway http://microbialgenome.org/

Publications and Presentations

GTL Science and Administration

- Program Overview, Jan 2003
  http://DOEGenomesToLife.org/pubs/overview_screen.pdf
- GTL draft facilities strategy and plan submitted to the Biological and Environmental Advisory Committee by the Life Sciences Division of the Biological and Environmental Research program for the Dec 3-4, 2002 meeting
  http://DOEGenomesToLife.org/pubs/GTLFac34BERAC45.pdf
- Resources and Technology Centers for Biological Discovery in the 21st Century brochure (Published version with images derived from working paper below; June 2002)
  http://DOEGenomesToLife.org/research/GTLFacilities18screen.pdf
- Working paper presented to the BERAC; April 25, 2002
  http://DOEGenomesToLife.org/pubs/gtl_facilities.pdf
- Primer Pictorial; Oct 2001
  http://DOEGenomesToLife.org/primer.html
- Genomes to Life Roadmap; April 2001
  http://DOEGenomesToLife.org/roadmap/index.html
- Bringing the Genome to Life Report of a subcommittee of the Biological and Environmental Research Advisory Committee (BERAC); August 2000
  http://DOEGenomesToLife.org/history/genome-to-life-rpt.html

Workshop Reports on Computing

- Mathematics for GTL Workshop, Gaithersburg, Maryland; March 18-19, 2002
  http://DOEGenomesToLife.org/pubs/GTLMath-6.pdf
- Computer Science for GTL Workshop, Gaithersburg, Maryland; March 6-7, 2002
  http://doegenomestolife.org/pubs/computerscience/
- Computing Infrastructure and Networking Workshop, Gaithersburg, Maryland; Jan 22-23, 2002
  http://DOEGenomesToLife.org/compbio/mtg_1_22_02/infrastructure.html
- Visions for Computational and Systems Biology Workshop for the Genomes to Life Program; September 6-7, 2001 (Notes with Executive Summary)
  http://DOEGenomesToLife.org/pubs/CompbioVisions-4.pdf
- First GTL Computational Biology Workshop; Aug 7-8, 2001
  http://DOEGenomesToLife.org/compbio/mtg8_8_01/titlepage.htm

Workshop Reports on Technologies

- Report on the Imaging Workshop for the Genomes to Life Program, April 16-18, 2002
  http://DOEGenomesToLife.org/technology/imaging
- Executive Summary of Workshop on Imaging Technologies, April 16-17, 2002
  http://DOEGenomesToLife.org/pubs/imaging_summary.pdf
- Imaging Technology Additional Reading
  http://DOEGenomesToLife.org/technology/imaging/workshop2002/biblio.html
- Genomes to Life: Technology Assessment for Mass Spectrometry, Dec 10-11, 2001
  http://DOEGenomesToLife.org/pubs/mass_spec.pdf

Energy Security and Global Climate Change

- "Energy Security and Climate Stabilization brochure"; August 2002
  http://DOEGenomesToLife.org/energy_carbon.pdf
- "Energy and Climate Change Payoffs poster presented at the G8 Energy Conference"; April 2002
  http://DOEGenomesToLife.org/warming/climateposter.pdf
- James Edmonds, Nov 27, 2001 to Biological and Environmental Research Advisory Committee
  http://DOEGenomesToLife.org/nov27slides/edmonds.htm
- Robin Graham, Nov 27, 2001 to Biological and Environmental Research Advisory Committee
  http://DOEGenomesToLife.org/nov27slides/graham.htm
- Workshop on The Role of Biotechnology in Mitigating Greenhouse Gas Concentrations; June 23, 2001: A Workshop Summary by Ken Nealson and J. Craig Venter, Workshop Cochairs
  http://DOEGenomesToLife.org/warming/GTLBiotechwksp.htm

Bioremediation

- "Innovative Approaches for Cleaning Up and Treating Hazardous Wastes at DOE Sites" brochure; June 2002
  http://DOEGenomesToLife.org/cleanup.pdf
- Presentation by Blaine Metting on November 27, 2001 to Biological and Environmental Research Advisory Committee
  http://DOEGenomesToLife.org/nov27slides/metting.htm

Video

- 3-D image of a *Shewanella oneidensis* cell
  http://DOEGenomesToLife.org/pubs/video.html#shewanella

Related Publications

- Microbial Ecology and Genomics: A Crossroads of Opportunity. A Report from the American Academy of Microbiology" (based on a colloquium held Feb 23-25, 2001)
  http://www.asmusa.org/acasrc/pdfs/MicroEcoreport.pdf
- "A Target Area of the Scientific Simulation Initiative"
  http://cbcg.lbl.gov/ssi-csb/HomeDetail.html
- "Computational Challenges in Structure and Functional Genomics" (*IBM Systems Journal*, Vol 40, No 2, 2001)
  http://www.research.ibm.com/journal/sj/402/headgordon.pdf
- "The Biomedical Information Science and Technology Initiative" (June 3, 1999)
  http://www.nih.gov/about/director/060399.htm
- "Advanced Computational Structural Genomics"
  http://cbcg.lbl.gov/ssi-csb/Meso.html
- Presentation by Gary Johnson, Oct 26, 2001 to the Office of Advanced Scientific Computing Research Advisory Committee
  http://DOEGenomesToLife.org/compbio/johnson.htm

Author Index

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z

Institution Index

S

U

W

Y

AGENDA
Genomes to Life Contractor-Grantee Workshop I
February 9–12, 2003
Sheraton National Hotel, Arlington, Virginia

Sunday, February 9

7:00–9:00 p.m.	No-host mixer, 16th floor, Galaxy Ballroom

Monday, February 10

8:30 a.m.	Welcome, logistics, program overview, introduction to breakout sessions
9:30	George Church, Harvard Univ.
10:00	Michelle Buchanan, ORNL
10:30	BREAK
11:00	Grant Heffelfinger, Sandia National Lab.
11:30	Derek Lovley, Univ. of Mass. at Amherst
12:00	Adam Arkin, LBNL
12:30–2:00	LUNCH
2:00–4:00	Session A Breakouts
4:00–4:30	BREAK
4:30–6:30	Poster Session A. Poster boards at the meeting will be numbered; poster placement will correspond to poster number as indicated on abstract listing.

Tuesday, February 11

8:30 a.m.	Lucy Shapiro, Stanford Univ.
9:00	Carrie Harwood, Univ. of Iowa
9:30	Bernhard Palsson, Univ. of Ca., San Diego
10:00	BREAK
10:30	Town Hall – Data Sharing
11:30	Production Genomics Facility – DNA sequencing capabilities, process, issues
12:00	Related programs at other federal agencies:
	NIGMS (NIH), Jerry Li; NSF, Matt Kane; DARPA, Sri Kumar
12:30	LUNCH
2:00–4:30	Session B Breakouts
4:30–6:30	Poster Session B. Poster boards at the meeting will be numbered; poster placement will correspond to poster number as indicated on abstract listing.

Wednesday, February 12

8:30–10:00 a.m.	*Shewanella* Federation
8:30–8:45	*Shewanella* Federation Overview – Jim Fredrickson, PNNL
8:40–8:55	Cultivation Techniques & Initial Collaborative Experiments – Yuri Gorby, PNNL
8:55–9:10	Microarray Analyses – Jizhong Zhou, ORNL
9:10–9:30	Proteome Analyses Accurate Mass Tags – Richard Smith, PNNL 2D Gels – Carol Giometti, ANL
9:30–9:45	Data Analysis – Eugene Kolker, BIATECH
9:45–10:00	Future Directions using Continuous Cultivation – Ken Nealson, USC
10:00	BREAK
10:30–12:00	Town Hall – GTL Facilities Discussion – Marvin Frazier, DOE
12:00–12:45	Blue Skies Forum – Craig Venter, Institute for Biological Energy Alternatives
12:45	Adjourn

Breakout Sessions

Breakout sessions will focus on key topics of interest to the Genomes to Life program. They are intended to be open discussion sessions for all interested participants. The sessions will begin with several short presentations to help stimulate discussion and will be led by several co-chairs. Sessions may focus on several key questions or challenges.

1. Data standards, data sharing, software openness:
   Session A Co-chairs – Ed Uberbacher, ORNL; Mike Colvin, LLNL; Arie Shoshani, LBNL
2. Environmental genomics / microbial communities, including computational needs:
   Session A Co-chairs – Derek Lovley, Univ. of Massachusetts, Amherst; Barbara Methe, TIGR
3. Proteomics, molecular machines, including computational needs:
   Session A Co-chairs – Michelle Buchanan, ORNL; Grant Heffelfinger, Sandia
4. New insights from comparative genomics:
   Session B Co-chairs – Jim Fredrickson, PNNL; Susan Lucas, DOE Joint Genome Institute
5. Technology development / integration:
   Session B Chair – George Church, Harvard Univ.
6. Regulatory networks, including computational needs:
   Session B Co-chairs – Adam Arkin, LBNL; Bob Tabita, Ohio State Univ.

U.S. Department of Energy
Office of Biological and Environmental Research (SC-72)
Office of Advanced Scientific Computing Research (SC-30)
Germantown Building
1000 Independence Ave., SW
Washington, DC 20585-1290