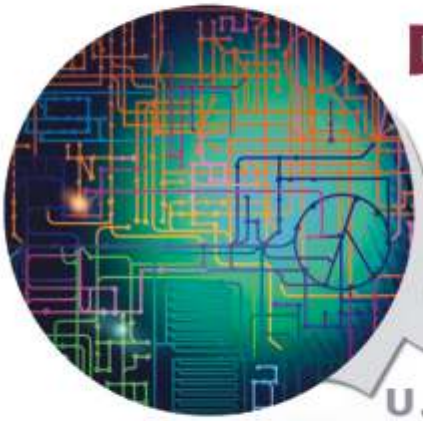
DOE/SC-0081



DOE GENOMICS:GTL

ACCELERATING DISCOVERY FOR ENERGY AND ENVIRONMENT

OFFICE OF SCIENCE
U.S. DEPARTMENT OF ENERGY

Contractor-Grantee Workshop II

Washington, D.C.

February 29 – March 2, 2004

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

Office of Science
U.S. DEPARTMENT OF ENERGY

Genomics:GTL 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

For print copies, 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 29, 2004,
at the Genomics:GTL Web site: http://doegenomestolife.org/pubs/2004abstracts/

Abstracts for this publication were submitted via the Web.

DOE/SC-0081

DOE GENOMICS:GTL
ACCELERATING DISCOVERY FOR ENERGY AND ENVIRONMENT
OFFICE OF SCIENCE
U.S. DEPARTMENT OF ENERGY

Contractor-Grantee Workshop II

Washington, D.C.

February 29–March 2, 2004

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
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

Posters are indicated by the large numbers.

Oak Ridge National Laboratory and Pacific Northwest National Laboratory

Sandia National Laboratories

Posters are indicated by the large numbers.

University of Massachusetts, Amherst

Posters are indicated by the large numbers.

Shewanella Federation

Institute for Biological Energy Alternatives

Posters are indicated by the large numbers.

Posters are indicated by the large numbers.

Posters are indicated by the large numbers.

Posters are indicated by the large numbers.

Posters are indicated by the large numbers.

Posters are indicated by the large numbers.

Posters are indicated by the large numbers.

Protein Production and Molecular Tags

Posters are indicated by the large numbers.

Proteomics

Metabolomics

Posters are indicated by the large numbers.

Posters are indicated by the large numbers.

Welcome to Genomics:GTL Contractor-Grantee Workshop II

Welcome to the second of what we hope will be many Genomics:GTL (formerly Genomes to Life) contractor-grantee workshops. Although only in its third 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 the leading edge of a new era in biology in which we will continue to use a broad array of innovative technologies and computational tools to systematically leverage the knowledge and capabilities brought to us by DNA sequencing projects. With these resources, we will seek to understand 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 you to develop technologies that enable us to "see" biology happen at finer scales of resolution. It also will require substantial integration of our broad capabilities in mathematics and computation with our new knowledge of biology. Thus, the look of the scientists and research projects at this second GTL workshop is considerably different from that of last year's workshop. 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.

Microbes remain GTL's principal biological focus. In the complex "simplicity" of microbes—both individual microbes and complex microbial communities—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 usable in all areas of biological research.

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

This workshop also kicks off a year-long process to develop a GTL roadmap that will outline the Genomics:GTL program:

- Research plan and science deliverables;
- Infrastructure, including national user facilities;
- Impacts on energy production, environmental cleanup, and carbon sequestration;
- Ethical, legal, and societal considerations;
- Interagency partnerships and coordination; and
- Program governance.

We will be calling on many of you over the coming weeks and months to provide critical input to and review of this roadmap. Our plan is to have the document ready for National Academies review by January 2005. We thank you in advance for your help in developing this important document that will help guide the direction of and serve as an informational resource on the GTL program.

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

Ed Oliver
Associate Director of Science for
Advanced Scientific Computing Research
Office of Science
U.S. Department of Energy
ed.oliver@science.doe.gov

Genomics:GTL Program Projects

Harvard Medical School

Microbial Ecology, Proteogenomics, and Computational Optima

1

Flux Balance Based Whole-Cell Modeling of the Marine Cyanobacterium *Prochlorococcus*

George M. Church[1] (g1m1c1@arep.med.harvard.edu), Daniel Segre[1], Xiaoxia Lin[1], Kyriacos Leptos[1], Jeremy Zucker[2], Aaron Brandes[2], Dat Nguyen[1], and Jay MacPhee[1]

Department of Genetics, Harvard Medical School, Boston, MA and [2]Dana-Farber Cancer Institute, Boston, MA

http://arep.med.harvard.edu/DOEGTL/

The marine unicellular cyanobacterium *Prochlorococcus* is the dominant oxygenic phototroph in the tropical and subtropical oceans, and contributes to a significant fraction of the global photosynthesis (Rocap et al, 2003). Our goal in this project is to develop whole-cell mathematical models for studying the metabolism of this cyanobacterium using flux balance based approaches, which has proven very successful in performing whole-cell modeling for a variety of microorganisms (Price et al, 2003). An especially interesting challenge is the inclusion of photosynthesis pathway in our model. Night-day cycles are known to play a central role in the metabolism of *Prochlrococcus*, and different strains are adapted to different light intensities and wavelengths. Flux balance models give the opportunity to study quantitatively the influence of photon fluxes on global cell behavior.

The completion of the *Prochlorococcus* genome sequencing has provided us a promising starting point for building whole-cell flux balance models of this bacterium. By utilizing an automatic bioinformatics pipeline which was recently developed (Segre et al, 2003), we have combined the genome annotation of *Prochlorococcus* MED4, a high-light-adapted strain, with an extensive pathway/genome database, MetaCyc (Karp et al, 2002), and generated a *Prochlorococcus* MED4 pathway database. This organism-specific pathway database is then used to generate flux balance models in which given the stoichiometric matrix representing the metabolic networks and limits on nutrient uptakes, linear programming (LP) or other optimization techniques are used to calculate the flux distribution that reflects the metabolic state of the cell. Our preliminary studies have shown that a substantial number of the biomass components can not be produced with the current identified metabolic networks. This is mainly due to i) incomplete annotation of the genome, for example, not identifying a gene encoding the enzyme catalyzing a metabolic reaction in the biosynthesis pathway of a certain amino acid; and ii) incomplete inclusion of pathways from the

MetaCyc database. In order to generate flux balance models that can capture the primary components of the metabolic networks of *Prochlorococcus* and then can be used to study its genotype-metabolic phenotype relationship under varying conditions, we are currently improving and refining the models by i) using network debugging methods to identify missing reactions/pathways in the constructed *in silico* metabolic network; ii) including additional reactions/pathways based on information from a variety of other sources, such as identification of enzymes through manual search of homologs, proteomic data, existing knowledge about the bacterium's metabolism, etc.

Another important requirement for the construction of whole-cell flux balance models of *Prochlorococcus* is to incorporate an appropriate set of transport reactions, which are currently lacking in the MetaCyc database. Approximately 50 transport proteins have been classified according to the Transport Classification system, which includes substrate specificity, through a combination of TC-BLAST, pfam, COG, and phylogenetic tree analysis (available at http://membranetransport.org). The transport reactions associated with these proteins can be deduced directly from their Transport classification number. We are working closely with the curators of MetaCyc and the Membrane transport database to incorporate these reactions into the pathway/genome database for *Prochlorococcus*.

Upon the successful construction of whole-cell flux balance models for *Prochlorococcus*, we plan to i) investigate how the metabolic network of this cyanobacterium works to enable it grow/live under its natural environmental conditions, in specific, in the light and in the dark; ii) investigate the differences between high-light-adapted strains, for example, MED4, and low-light-adapted strains, for example, MIT9313, by comparing the structures of their metabolic networks and the calculated flux distributions under varying conditions; and iii) investigate the effect of gene knockouts on cellular properties, such as growth rate and photosynthesis, using the MOMA approach developed earlier in the Church lab (Segre et al, 2002). Hypotheses generated with flux balance models will be tested experimentally using expression and proteomic data.

Reference

1. Karp PD, Riley M, Paley S, and Pellegrini-Toole A (2002) The MetaCyc Database. *Nucleic Acids Research* **30**(1):59-61.

2. Price ND, Papin JA, Schilling CH, and Palsson BO (2003) Genome-scale microbial in silico models: the constraints-based approach. *Trends Biotechnol*. **21**(4): 162-169.

3. Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, Johnson ZI, Land M, Lindell D, Post AF, Regala W, Shah M, Shaw SL, Steglich C, Sullivan MB, Ting CS, Tolonen A, Webb EA, Zinser ER, and Chisholm SW (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**(6952):1042-1047.

4. Segre D, Vitkup D, and Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc. Nat. Acad. Sci USA* **99**: 15112-7.

5. Segre D, Zucker J, Katz J, Lin X, D'haeseleer P, Rindone W, Karchenko P, Nguyen D, Wright M, and Church GM (2003) From annotated genomes to metabolic flux models and kinetic parameter fitting. *Omics* 7:301-16.

Lawrence Berkeley National Laboratory

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

2

VIMSS Computational Microbiology Core Research on Comparative and Functional Genomics

Adam Arkin[1,2,3] (aparkin@lbl.gov), Eric Alm[1], Inna Dubchak[1], Mikhail Gelfand[4], Katherine Huang[1], Kevin Keck[1], Frank Olken[1], Vijaya Natarajan[1], Morgan Price[1], and Yue Wang[2]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]University of California, Berkeley, CA; [3]Howard Hughes Medical Institute, Chevy Chase, MD; and [4]Research Institute for the Genetics and Selection of Industrial Microorganisms, Moscow, Russa

The primary roles of the Computational Core are to curate, analyze, and ultimately build models of the data generated by the Functional Genomics and Applied Environmental Microbiology Core groups. The near-term focus of the computational group has been to build the scientific and technical infrastructure necessary to carry out these roles. In particular, the efforts of the computational group have been directed toward three objectives: genomics and comparative genomics, curation and analysis of experimental data from the other core groups, and modeling. Central to each of these goals has been the development of a comprehensive relational database that integrates genomic data and analyses together with data obtained from experiment.

VIMSS DB. At present, well over 100 microbial genomes have been sequenced, and hundreds more are currently in the pipeline. Despite this fact, tools to explore this wealth of information have focused on individual genome sequences. The VIMSS Comparative Genomics database and web-based tools are designed to facilitate cross-species comparison, as well as to integrate experimental data sets with genome-scale functional annotations such as operon and regulon predictions, metabolic maps, and gene annotations according to the Gene Ontology. Over 130 complete genome sequences are represented in the VIMSS Comparative Genomics Database, which is implemented as a MySQL relational database, a Perl library for accessing the database, and a user-friendly website designed for laboratory biologists (http://escalante.lbl.gov). This database is currently being augmented with a novel graph for the efficient query of biological pathways and supporting data. A generic java-based tool for the graphical construction of queries on representations of relational database schema (particular for pathways) in nearly finished and will be applied to VIMSS DB in first quarter 2004.

Web-Based Tools. The VIMSS Comparative Genome Browser allows users to align any number of genomes and identifies predicted orthology relationships between genes. Users can save genes of interest for use in the VIMSS Bioinformatics Workbench (VBW), explore individual genes in depth for information about sequence domains, BLAST alignments, predicted operon structure and functionally related genes inferred from a combination of comparative genomics

methods and microarray experiments. The VertiGO comparative gene ontology browser allows users to simultaneously view the genetic complement of any number of genomes according to the Gene Ontology hierarchy. A metabolism browser based on the KEGG metabolic maps allows browsing either the set of enzymes predicted to be present in a single genome, or a comparison highlighting the metabolic differences between two genomes. VBW allows users to create and save lists of genes of interest, and use these lists to investigate phylogenetic relationships by making multiple sequence alignments and phylogenetic trees, as well as apply DNA motif-finding software to identify potential regulatory elements in upstream sequences. Novel motif finding algorithms exploiting the comparative analysis of orthologous proteins have already been accurately difficult motifs such as those from the merR family of regulators of heavy-metal resistance.

Genome Annotation. One of the stated goals of the GTL program is to produce next-generation annotation of target genomes including automated gene functional annotations and prediction of gene regulatory features along with validation of these in silico methods. The most fundamental unit of gene regulation in bacteria is the operon, which is a set of genes that are cotranscribed on a single RNA transcript. Because few operons have been characterized experimentally outside the model organisms *E. coli* and *B. subtilis*, in silico operon prediction methods have been validated only in these two organisms. We have therefore made accurate and unbiased operon predictions in all bacteria a priority for the computational group. To avoid bias that might arise from using experimental data from only two organisms, we have opted to avoid the use of experimental data entirely using techniques from the field of unsupervised machine learning, and we used gene expression data to estimate the accuracy of our predictions. Key to the success of this approach has been integrating experimental data from the Functional Genomic Core group into our Comparative Genomics Database to validate our in silico procedures. Using our operon prediction tool, we have established that, contrary to reports in the literature, the bacterium *Helicobacter pylori* has a large number of operons. In addition, by examining unusually large non-coding regions within highly conserved operons, we have identified putative pseudogenes in *Bacillus anthracis* that allow us to make phenotypic predictions about the motility of the sequenced Ames strain. As a critical test of our automated genome annotations, we are hosting a genome annotation jamboree in April at the Joint Genome Institute, in which our automated predictions will be verified by human curators. We expect that our annotations, along with confidence levels, will reduce the manual curation workload allowing participants to focus most of their efforts on scientific hypothesis testing.

Functional Genomics. The Functional Genomics Core group is beginning to produce large data sets detailing the response of our target organisms to a variety of stress conditions. The Computational Core group is charged with the responsibility to: store and redistribute these data; assist in the statistical analysis and processing of raw data; and to facilitate comparison of experiments performed with different experimental techniques, different conditions, or different target organisms. As a test case, we have focused most of our efforts in this direction toward gene expression microarray experiments. Among the challenges in the representation of microarray data is developing a data schema that includes both raw and processed data, metadata describing the experimental conditions, and a technical description mapping, for example, each array spot to a corresponding region of the genome sequence and to the set of annotated genes (and their orthologs in other species). We are actively following the development of standards for the representation of this type of data (see Data Management below), and in the meantime have implemented our own simple formats aimed at quick integration with our Comparative Genomics Database. To interpret the results of these experiments, it was necessary

to develop a standard set of procedures for data normalization and significance testing and apply it uniformly to raw data from each experiment set, as processed data from different labs commonly involve slightly different analytical techniques. By establishing common methodologies, and a common repository for different experimental results, we were able to meet the goal of facilitating comparative studies as well as using the functional genomic data to test hypotheses generated from our comparative genomic analysis. The methods have been applied to the analysis of pH, salt and heat stress data from *Shewanella oneidensis*. Results from this analysis will be described.

Data Management. During the first year of the project, laboratories in the project began putting in place experimental procedures and are now beginning to produce substantial amounts of data. There is a critical need to define what descriptions of data and experimental procedures (protocols) and factors need to be developed and captured, and to put in place procedures for documenting and recording that information. Recognizing this need, we are in the process of reviewing how experimental procedures are being documented and how experimental factors are being recorded by LBNL affiliated laboratories. This information will be used not only to facilitate information and data acquisition procedures, but also to to enhance and upgrade the BioFiles system for data uploading and the underlying database management system. Working with a consortium of researchers from the wider GTL community we have produced a report on the current status of National Data standards and their advantages and deficiencies and produced a plan for developing standardization of metadata and data representation.

3

Managing the GTL Project at Lawrence Berkeley National Laboratory

Nancy A. Slater (naslater@lbl.gov)

Lawrence Berkeley National Laboratory, Berkeley, CA

The effective management of the GTL systems biology project at Lawrence Berkeley National Laboratory (LBNL) is essential to the success of the project. The comprehensive management plan for the project includes milestone planning and project integration, a plan for communicating and collaborating with the project stakeholders, financial management and website updates. In addition, the management plan incorporates reviews by committees, including a monthly Executive Committee review comprised of LBNL leadership, an annual Scientific Advisory Committee review, a biannual Technical Advisory Panel review to ensure that the project's technical development is aligned with related DOE efforts, and a monthly Steering Committee conference call where the project leaders discuss the project's progress and status.

A key responsibility in the project management process is troubleshooting problems related to the scientific and financial management of the project. There is a delicate balance between having adequate resources to achieve the scientific objectives of the project and working within the funding levels of the project. If an area is falling behind on achieving their scientific milestones, the project manager must work closely with the researchers to resolve problems as efficiently and effectively as possible.

Milestone Planning and Project Integration

A detailed list of project deliverables and milestones is updated by the PIs at the beginning of each fiscal year. The process of updating and reviewing milestones ensures that the goals of each PI are aligned with the overall goals of the project. These milestones are the basis for an integrated project schedule, which is managed using Microsoft® Project. The project schedule is updated monthly, and progress is reported through progress reports and teleconferences with the PIs. The updated project schedule is posted to the project website, so that all of the collaborators have access to the most recent status of the project.

The project is divided into three separate Core groups, and the integration plan for the project assures that the Core groups work together toward the objectives of the project. The Core Research group leaders are responsible for ensuring smooth operation of their section of the project as well as cooperation with the other groups. For example, the Applied Environmental Microbiology leader is responsible for ensuring that cell culture protocols are acceptable to the Functional Genomics Core, who will ultimately use the cell cultures for experiments. The Functional Genomics Core leader is responsible for ensuring quality control for data production and timely data uploads into the database. The Computational Core leader is responsible for ensuring that data entry, querying, and curation interfaces serve the needs of the other groups, and that the models are useable to biologists outside of the modeling group. The success of each group is interdependent on a well-integrated project team.

Communication and Collaboration

The GTL project at LBNL is a collaborative effort between seven institutions, thirteen researchers and their associated laboratories. The project's communications plan consists of a variety of media, including a project website, monthly group meetings, conference calls, an annual retreat, workshops at conferences, and monthly progress reports.

The monthly group meetings include a presentation from one of the Core Research groups, and it is attended by the local, northern California GTL project team members. There are a several conference calls that are held on a regular basis, including a monthly Steering Committee meeting in which all of the researchers participate, a monthly BioFiles conference call in which a representative from each laboratory discusses data generation, uploads and handling, and a quarterly conference call with DOE. The LBNL project has an annual retreat in which the researchers present data and findings related to their area of focus and other laboratory team members (Computer Science Engineers, Microbiologists, Database Managers, Graduate Students, Post Docs, etc.) present posters in a poster forum. The annual retreat has proven to be very successful in building working relationships among the dispersed group. The LBNL GTL project will be participating in several workshops at international conferences in 2004. The monthly progress reports are comprised of input from each researcher, and include updates regarding the status of the milestones, planned work and problems/issues that they encountered.

Website Updates

The GTL project at LBNL is the inaugural project for the Virtual Institute of Microbial Stress and Survival (VIMSS), and details regarding the project are located on the world wide web at http://vimss.lbl.gov. This website serves as a tool for communicating the status of the project as well as:

- an overview of the GTL project at Berkeley and links to the key personnel working on the project
- a link to Comparative Genomics Tools such as the Comparative Genome Database, the Genome Browser, and Operon and Regulon Prediction tools
- a link to the BioFiles repository of project data
- a discussion board for project team members to interact and post protocols, questions and solutions
- a job board with available GTL-related positions
- a calendar of upcoming meetings and events

Financial Management

Each of the researchers provides input into the annual spend plan for the project. The finances of the project are tracked on a continuous basis, and the researchers receive monthly reports showing actual costs verses the spend plan. The finances of the project are maintained using software packages at LBNL as well as spreadsheets and charts. These tools allow the Project Manager to identify spending trends, so that appropriate can be taken to keep the project aligned with the annual spend plan. The Executive Committee reviews the project financial reports monthly.

4

VIMSS Applied Environmental Microbiology Core Research on Stress Response Pathways in Metal-Reducers

Terry C. Hazen*[1] (TCHazen@lbl.gov), **Hoi-Ying Holman**[1], Sharon E. Borglin[1], Dominque Joyner[1], Rick Huang[1], Jenny Lin[1], **David Stahl**[2], Sergey M. Stolyar[2], **Matthew Fields**[3], **Dorothea Thompson**[3], **Jizhong Zhou**[3], **Judy Wall**[4], H.-C. Yen[4], and **Martin Keller**[5]

*Presenting author

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]University of Washington, Seattle, WA; [3]Oak Ridge National Laboratory, Oak Ridge, TN; [4]University of Missouri, Columbia, MO; and [5]Diversa Corporation, San Diego, CA

Field Studies

Sulfate-reducing bacteria.: Sediment samples from different depths at the NABIR Field Research Center in the background, Areas 1, 2, and 3 sites have been used for the enrichment of sulfate-reducing microorganisms. Sulfate-reducing enrichments have been positive for sediments in Areas 1 and 2 when lactate or acetate were used as electron donors, and some of the enrichments differ in the capacity to reduce cobalt, chromium, and uranium. Groundwater enrichments from Areas 1, 2, and 3 all displayed sulfate-reduction with different electron donors (lactate, butyrate, acetate, pyruvate) and these enrichments could also reduce iron, cobalt, and chromium. Subsurface sediments from the wells FWB-107 (13.2 m) and FWB-109 (15.4 m) in Area 3 were serially diluted in a basal salts medium that contained lactate and ethanol with different electron acceptors. The results suggested that in the sampled sediments (13 to 15 m) nitrate-reducers were approximately 3500 to 5400 cells/g, iron-reducers 50 to 1700 cells/g, and sulfate-reducers 240 to 1100 cells/g. The predominant population (25%) of the 10-2 sulfate-reducing dilution had 88%

sequence identity with *Desulfosporosinus blif*. Subpopulations that had 95% to 97% sequence identity with *Desulfosporosinus orientis* constituted for an additional 37% of the library. Other clones had 98% sequence identity with *Clostridium chromoreductans*.

Clone libraries. Since stress response pathways are clustered on chromosomal DNA fragments and generally vary in length from 20-40 kb, it is essential to clone large DNA fragments to capture entire pathways. We have developed effective DNA extraction methods and vector/host systems that allow stable propagation of large DNA fragments in *E. coli*. Processed environmental samples are embedded in agarose noodles for protein digestion and release of high molecular weight DNA. In stressed environments, organism concentrations are often very low, so we have developed a method for increasing the concentration of large DNA by amplification with a phage polymerase. After amplification, the DNA is partially digested with restriction enzymes, and size-selected by agarose gel electrophoresis. It is then ligated to fosmid arms and packaged into phage lambda particles that are used to infect *E. coli*. The microbial diversity of the libraries is determined with Terminal Restriction Fragment Polymorphism (T-RFLP). Large fragment DNA has been extracted and amplified from 15 NABIR FRC samples (comprising 3 areas at various depths). Small insert DNA libraries have been constructed from most of these samples, and large insert DNA libraries are in various stages of construction. T-RFLP and DNA sequencing are being used to quality control the resulting libraries.

Enrichments. Seven *Desulfovibrio* strains were isolated from lactate-sulfate enrichment of sediment taken from the most contaminated region of Lake DePue, IL. Their 16S rRNA and dsrAB genes were amplified and sequenced. They all were identical to each other and virtually identical to the corresponding genes from *D. vulgaris* Hildenborough. One mismatch was observed in the16S rRNA gene and one in dsrAB. Different fragment patterns confirmed that the DePue isolates were similar but not identical to *D. vulgaris* Hildenborough. Pulse field electrophoretic analysis of I-CeuI digests revealed that both isolates had five rRNA clusters, the same as *D. vulgaris* Hildenborough. However, the length of one chromosomal segment in the DP isolates was considerably shorter than the corresponding fragment from *D. vulgaris* Hildenborough, suggesting the presence of a large deletion in the genomes of the isolates (or insertion in *D. vulgaris* Hildenborough).

Culture and Biomass Production

Defined Media – Growth. A defined medium for optimal growth and maximum reproducibility of *Desulfovibrio vulgaris* was developed for biomass production for stress response studies. The medium was optimized by evaluating a variety of chemical components, including the removal of yeast extract, excess sulfate, and Fe, and redox conditions to optimize cell density and generation times, and to reduce lag times. Growth was monitored using direct cell counts, optical density, and protein concentration. The generation time for *D. vulgaris* in the original Baar's medium was 3 h, reaching a maximum density of 10^8 cells/ml and 0.4 $OD_{600\ nm}$. The generation time for *D. vulgaris* on LS4D was 5 h, with a maximum cell density of 10^9 cells/ml and a 0.9-1.0 $OD_{600\ nm}$. LS4D is well suited for the monitoring protocols, as well as the equipment and large scale processing needed for biomass production.

Dual culture systems. Co-cultures of two different *Desulfovibrio* species (*Desulfovibrio vulgaris* Hildenborough and *Desulfovibrio* sp.PT2) syntrophically coupled to a hydrogenotrophic methanogen (*Methanococcus maripaludis*) on a lactate medium without sulfate has been established and characterized. No appreciable

growth was observed in 50 mM lactate for single-organism cultures. Following optimization of the ionic composition (MgCl2 and NaCl) of the medium, stable co-cultures were established having generation times of 25h-1 and 35 h-1 for *D. vulgaris* and *Desulfovibrio* sp. PT2 co-cultures respectively. Both co-cultures degraded lactate to acetate, methane, and carbon dioxide. No other organic acids were detected during the course of experiments. Approximately 1mol of acetate and 1mol of methane was produced from two mole of lactate by both co-cultures during most active period of growth. The stability of established methanogen-SRBs co-cultures (*Desulfovibrio vulgaris* or *Desulfovibrio* sp. PT2 with *M. maripulidis*) was confirmed by serial transfer (six times).

Biofilm reactors. Initial characterization of *Desulfovibrio vulgaris* growth as a biofilm was evaluated using a 600ml biofilm reactor containing 3mm glass beads as growth substratum and the B3 culture medium (16mM lactate and 28 mM sulfate). The ratio of flow rates through an internal recirculation loop to influent was maintained at 100:1, evaluating two different influent flow rates (0.5ml/min or 30ml/hr). Formation of a loose biofilm was associated with significant gas accumulation within the reactor. The system in now being modified to incorporate a gas trap in the re-circulation loop.

FairMenTec (FMT) chemostat. A pilot run with *Desulfovibrio vulgaris* Hildenborough in the FMT bioreactor in chemostat mode was completed. The bioreactor was operated using the LS4D medium with 45mM lactate, 50 mM sulfate, and Ti-citrate at 1/3 standard formulation (subsequent batch cultures have shown improved growth with further reduction of the Ti-citrate to 1/6 standard formulation). Varying flow rates and medium compositions were evaluated.

Oxygen Stress Experiments

Protocols. Since episodic exposure to air or oxygenated ground water is common at contaminated sites, we decided to focus on oxygen stress of *D. vulgaris* for our initial studies. To accommodate the all the investigations that would require simultaneous harvesting of biomass for studies on proteomics, transcriptomics, metabolomics and phenotypic studies a batch culture system was developed for 2000 ml cultures that could be sparged with nitrogen or air to control stress in water baths using rigorous quality control on culture age, sampling, defined media, chain of custody, and harvesting times and techniques.

Phenotypic responses. Desulfovibrio vulgaris enters a new phenotypic state when confronted with a sudden influx of oxygen. Using SEM and TEM microscopy we observed that during the first 24-72 h of exposure to air *D. vulgaris* cells are negatively aerotactic, gradually they loose their flagella, and begin to elongate, by 20 days exposure they are 3-4 times larger and have a well developed exopolysaccharide sheath. At all times the cells were viable and recovered when put back under anaerobic conditions. Real-time analysis using Synchrotron Fourier Transform Infrared Spectromicroscopy enabled us to determine quantitative changes in peptides and saccharides in the living cells during exposure to air, thus providing the exact timing of cell changes in the stress response. During the early phase of the exposure, we observed decreases in total cellular proteins as well as changes in the secondary structures of proteins that are indicative of the changing of the local hydrogen-bonding environments and the presence of granular protein. During the late phase of the exposure, we observed the production of polysaccharides, concomitant with the production of the external sheath. The S-FTIR also demonstrated that the cells were viable within the sheath at 20 days exposure. Phospholipid fatty acid (PLFA) analysis confirmed that no biomass was loss during air sparging of stationary phase cells.

In addition, no change in the PLFA patterns were observed during air sparge, indicating neither cell growth nor death occurred. The PLFA extraction is being developed as a method for routine monitoring of cultures during biomass production and stress studies. Databases of lipid signatures of *D. vulgaris* during various growth conditions are being developed to augment the information produced from other VIMSS collaborators on proteomics and functional genomics.

5

VIMSS Functional Genomics Core: Analysis of Stress Response Pathways in Metal-Reducing Bacteria

Jay Keasling*[1] (keasling@socrates.berkeley.edu), Steven Brown[4], Swapnil Chhabra[2], Brett Emo[3], Weimin Gao[4], Sara Gaucher[2], Masood Hadi[2], Qiang He[4], Zhili He[4], Ting Li[4], Yongqing Liu[4], Vincent Martin[1], Aindrila Mukhopadhyay[1], Alyssa Redding[1], Joseph Ringbauer Jr.[3], Dawn Stanek[4], Jun Sun[5], Lianhong Sun[1], Jing Wei[5], Liyou Wu[4], Huei-Che Yen[3], Wen Yu[5], Grant Zane[3], **Matthew Fields**[4], **Martin Keller**[5] (mkeller@diversa.com), **Anup Singh**[2] (aksingh@sandia.gov), **Dorothea Thompson**[4], **Judy Wall**[3] (wallj@missouri.edu), and **Jizhong Zhou**[4] (zhouj@ornl.gov)

*Presenting author

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]Sandia National Laboratories, Livermore, CA; [3]University of Missouri, Columbia, MO; [4]Oak Ridge National Laboratory, Oak Ridge, TN; and [5]Diversa Corporation, San Diego, CA

Introduction: Environmental contamination by metals and radionuclides constitutes a serious problem in many ecosystems. Bioremediation schemes involving dissimilatory metal ion-reducing bacteria are attractive for their cost-effectiveness and limited physical detriment and disturbance on the environment. *Desulfovibrio vulgaris, Shewanella oneidensis*, and *Geobacter metallireducens* represent three different groups of organisms capable of metal and radionuclide reduction whose complete genome sequences were determined under the support of DOE-funded projects. Utilizing the available genome sequence information, we have focused our efforts on the experimental analysis of various stress response pathways in *D. vulgaris* Hildenborough using a repertoire of functional genomic tools and mutational analysis.

Transcript analysis: *D. vulgaris* is a δ-Proteobacteria with a genome size of approximately 3.6 Mb. Whole-genome microarrays of *D. vulgaris* were constructed using 70-mer oligonucleotides. All ORFs in the genome are represented with 3,471 (97.1%) unique probes and 103 (2.9%) non-specific probes that may have cross-hybridization with other ORFs. The microarrays were employed to investigate the global gene expression profiles of *D. vulgaris* in response to elevated salt and nitrite concentrations as well as exposure to oxygen. Approximately 370 ORFs were up-regulated (≥ 3-fold) and 140 ORFs were down-regulated when *D. vulgaris* cells were treated with 0.5 M NaCl for 0.5 hour. For example, genes involved in glycine, betaine, or proline transport were up-regulated 5-, 19- and 26-fold, respectively. Almost half of those genes with significant changes in expression are predicted as conserved hypothetical or hypothetical proteins. After 4-hour treatment, approximately 140 ORFs were up-regulated and more than 700 ORFs were down-regulated. Patterns of gene expression were distinctly different between time points. With 1 mM nitrite, *D. vulgaris* exhibited a lag phase of 28 h compared to a 5 h lag

phase in controls without nitrite addition. Strong nitrite treatment (5 or 10 mM) triggered a transient growth arrest and growth resumed gradually after 5 hours, suggesting the ability of *D. vulgaris* to overcome the toxicity of nitrite. Transcriptional profiling analysis was carried out following nitrite (10 mM) treatment. Transcripts highly up-regulated throughout the 5 h following nitrite shock included genes encoding two iron-sulfur cluster-binding proteins (65- and 15-fold) and a hybrid cluster (Fe/S) protein (24-fold). All three ORFs are annotated as redox-active proteins, and the hybrid cluster protein has been specifically proposed to participate in nitrogen metabolism. Surprisingly, the nitrite reductase genes were only moderately up-regulated (3-fold) as well as the formate dehydrogenase genes.

Protein analysis: A combination of Differential In-Gel Electrophoresis (DIGE), Isotope-Coded Affinity Tags (ICAT), and comprehensive proteome analyses were used to investigate the response of the *D. vulgaris* proteome to heat shock and O_2 stress. DIGE analysis of heat-shock stress response identified a total of 650 proteins. Sixty-three (63) proteins showed differences between the heat shocked (30 min) and control conditions. Using the complementary ICAT analysis we were able to identify a total of 219 proteins out of the *D. vulgaris* proteome. Out of this pool of proteins, 7 stress related proteins were identified. Similar analysis was also done with O_2-stressed cells. Based on cysteine containing tryptic peptides, a total of 92 proteins were identified. Among the identified proteins, 40 showed differences between the O_2-stressed and control conditions and of these at least 6 are known to be involved in O_2-stress response. Total comprehensive proteome analysis of *D. vulgaris* was also used to investigate differential protein expression induced by O_2-stress. Cellular tryptic-digested proteins from control and stressed cultures were analyzed by 3D μLC-MS-MS. A total of 1,791 unique proteins were identified.

Protein complex analysis: Based on the preliminary DIGE analysis of heat shock response in *D. vulgaris*, HSP70 (ORF00281) was identified as being involved in this stress condition. Western analysis using antibodies to the *E. coli* homolog (63% sequence identity) showed enhanced production of ORF00281 (Hsp70). The Anit-HSP70 antibody was then used to study bait-prey interactions in whole cell protein extracts from the heat shock condition using the Co-Immunoprecipitation kit for immobilization. Approximately 7 "pulled down" proteins bands were observed as possibly interacting proteins with HSP70. These bands were gel extracted and further analyzed by LC-MS-MS. To generate tagged proteins for identifying protein complexes in *D. vulgaris*, we have also explored the application of the IBA Strep-tag vector system for generating single chromosomal copies of genes fused to the tag sequence. We have generated a fusion of *dnaK* with the tag and have it integrated into the chromosome of *D. vulgaris* in single copy to determine the effectiveness of this system for providing complexes for proteomics analysis.

Metabolite analysis: We have developed a hydrophilic interaction chromatography method coupled to MS/MS detection to separate and identify nucleotides and redox cofactors. In addition, CE-MS methods were developed to analyze a variety of metabolites, including amino acids, nucleic acid bases, nucleosides, nucleotides, organic acid CoAs, redox cofactors, and the metabolic intermediates of glycolysis, the TCA cycle and the pentose phosphate pathway. All the methods were validated using *E. coli* cell extracts. Approximately 100 metabolites can be separated and identified. The development of an efficient method to obtain *D. vulgaris* metabolite extracts and its application to analyze stress responses in *D. vulgaris* are in progress.

Development of a genetic system: In efforts to improve the genetic versatility of *D. vulgaris*, spontaneous mutants resistant to either nalidixic acid or rifampicin were selected. These antibiotic resistances will allow counter-selection of sensitive *E. coli*

donors in conjugation experiments. Additional effort has been made to screen antibiotic sensitivity and resistance of *D. vulgaris*. The wild type was sensitive to G418 (400 μg/ml), ampicillin (20-50 μg/ml), carbinicillin (20-50 μg/ml) and resistant to gentamycin. The drug resistance markers present on many routinely used cloning vectors confer resistance to these antibiotics. Marker exchange mutagenesis of a number of regulatory genes is in progress by a procedure that will introduce molecular barcodes into the deletion sites. Sucrose sensitivity will be used to enrich for the second recombination event necessary to delete the wild-type copies of the target genes. Interestingly, we found that sucrose sensitivity is not expressed well in all *Desulfovibrio* strains. To further streamline methods for gene knockout, a vector system that uses a single cross-over event for gene deletion has been created. A 750-bp internal gene sequence flanked by 20 base pair UP and DOWN barcodes will used to simultaneously knock out and barcode each gene. Conjugal transfer using *E. coli* will be used to transform *D. vulgaris* with the suicide knockout vectors. The single cross-over gene deletion system also attempts to address issues of polar mutations. Additionally, a *lacZ* reporter will be incorporated into the site of gene deletion. Methylumbelliferyl β-D-galatoside, a fluorescent substrate the β-galactosidase reporter will be used for colony screening under anaerobic conditions. Finally, experiments to generate a library of transposon mutants are also underway. Putative mutants have been generated and will be screened for the presence and copy number of the transposon, stability of the antibiotic resistance, and randomness of the insertion.

Oak Ridge National Laboratory and Pacific Northwest National Laboratory

Genomics:GTL Center for Molecular and Cellular Systems

A Research Program for Identification and Characterization of Protein Complexes

6

Establishment of Protocols for the High Throughput Analysis of Protein Complexes at the Center for Molecular and Cellular Systems

Michelle V. Buchanan[1] (buchananmv@ornl.gov), Gordon Anderson[2], Robert L. Hettich[1], Brian Hooker[2], Gregory B. Hurst[1], Steve J. Kennel[1], Vladimir Kery[2], Frank Larimer[1], George Michaels[2], Dale A. Pelletier[1], Manesh B. Shah[1], Robert Siegel[2], Thomas Squier[2], and H. Steven Wiley[2]

[1]Oak Ridge National Laboratory, Oak Ridge, TN and [2]Pacific Northwest National Laboratory, Richland, WA

The first year of the Center for Molecular and Cellular Systems focused on evaluating methods for the efficient identification and characterization of protein complexes, identifying "bottlenecks" in the isolation and analysis processes, and developing approaches that could eliminate these bottlenecks. Oak Ridge National Laboratory (ORNL) and Pacific Northwest National Laboratory (PNNL) staff worked closely together to develop an integrated process for protein complex analysis. Emphasis has been placed on developing robust protocols that are adaptable to high throughput isolation and analysis methods. Progress has been made in all five major program areas—molecular biology, organism growth standardization, protein complex isolation/ purification, protein complex analysis, and bioinformatics/computation. During this first year, we have evaluated a two-phased approach to identify protein complexes. The first is an exogenous bait approach using one or more purified proteins to pull down the components of the associated protein complex. The second is an endogenous approach involving the *in vivo* expression of tagged proteins that are used to pull down the components of the associated protein complex. These complementary approaches each have their advantages. The first permits the high-throughput isolation of complexes from a single sample grown under defined conditions, while the latter permits the identification of complexes under cellular conditions, plus it can be combined with the development of new imaging methods to identify synthesis, turnover, and complex localization in real time. To test the established protocols two organisms were employed, *Rhodopseudomonas palustris* and *Shewanella oneidensis*. Techniques were optimized and standard protocols were established for endogenous complex isolation and exogenous complex isolation that will be deployed in year two of the project.

Considerable progress has also been made in advancing capabilities for the characterization of protein complexes that will minimize current bottlenecks, reduce the amount of sample required, and automate sample handling and processing. We have

made progress toward using an affinity-labeled crosslinker that allows selective isolation and subsequent mass spectrometric analysis of crosslinked peptides. Microfluidic technologies that reduce the amount of sample required for analysis and decrease the time required for separation have been applied to the analysis of peptides from protein complexes. Automated trypsinization and sample processing protocols have been developed that are designed around a 96-well format. Imaging of microbial cells, based upon introduction of fluorescent labels onto target proteins, has also been pursued. Automation of key parts of the cloning and complex isolation pipeline was initiated. Particular emphasis was given in this first year in establishing a common laboratory information management system (LIMS) and sample-tracking system that would facilitate distributed workflow across multiple laboratories. Results from the first year of this project have led to the design of a single, high-throughput production pipeline that will integrate efforts at both ORNL and PNNL. This will allow the high throughput analysis of hundreds of complexes during the next year. This pipeline will use complementary pull down methods, both endogenous and exogenous methods, to isolate protein complexes and provide greater confidence in complex characterization. This pipeline will be flexible to allow improved technologies to be incorporated as they are developed.

7

Isolation and Characterization of Protein Complexes from *Shewanella oneidensis* and *Rhodopseudomonas palustris*

Brian S. Hooker[1] (Brian.Hooker@pnl.gov), Robert L. Hettich[2], Gregory B. Hurst[2], Stephen J. Kennel[2], Patricia K. Lankford[2], Chiann-Tso Lin[1], Lye Meng Markillie[1], M. Uljana Mayer-Clumbridge[1], Dale A. Pelletier[2], Liang Shi[1], Thomas C. Squier[1], Michael B. Strader[2], and Nathan C. VerBerkmoes[2]

[1]Pacific Northwest National Laboratory, Richland, WA and [2]Oak Ridge National Laboratory, Oak Ridge, TN

As part of the Center for Molecular and Cellular Systems pilot project, we have been evaluating both endogenous and exogenous approaches for the robust isolation and identification of protein complexes. Exogenous isolation uses bait proteins to capture the protein complexes. To evaluate various exogenous isolation approaches, five complexes with differing physical characteristics were employed, both stable and transiently associating protein complexes. These complexes included RNA polymerase, the degradosome, and oxidoreductase, all stable protein complexes of varying complexity, and protein tyrosine phosphatase (Ptp) and methionine sulfoxide reductase (Msr), which are signaling proteins that form transient protein complexes. Evaluation of several different approaches has shown that covalent immobilization of the affinity reagent to a solid support works well to isolate the protein complex away from nonspecifically bound proteins, whether this involves direct bait attachment or the immobilization of an antibody against the bait or epitope tag. Approaches evaluated include covalent attachment of bait protein to glass beads that were subsequently used to capture protein complexes and expression of bait proteins with 6xhis tags, which were used to isolate complexes with nickel-chelating resins.

For endogenous complex isolation, we have developed a convenient, broad host range plasmid system to prepare tagged proteins in the native host. A series of expression vectors have been developed that can be used to transfect *E. coli* or *R. palustris*. These expression vectors have been constructed based on the broad host

range plasmid pBBR1MCS5. This vector was modified to contain the Gateway® pDEST multiple cloning region that allows site specific recombination cloning of targets from Gateway® entry plasmid. Four modified Gateway® destination vectors were constructed that can be used for expression of 6x histidine (6xhis) or glutathione(GST), N- or C- terminally tagged fusion proteins. Using this approach, methods have been developed to purify complexes using a double affinity approach (TAP) and complexes of suitable amounts and purity have been obtained for mass spectrometry evaluation. We have cloned a total of 22 *R. palustris* genes into these expression vectors to test expression and affinity purification methods for isolation of protein complexes using different affinity tags. The tested genes included those which code for proteins that are components of GroEL, GroES, ATP synthase, CO_2 fixation, uptake hydrogenase, ribosome, photosynthesis reaction center, Clp protease, and signal recognition. Results suggest while there was no one affinity tag which worked well for all genes tested, there was at least one fusion protein that expressed well for each targets tested. The 6xhis and V5 tag combination, does in fact yield a highly purified product in the test cases examined to date. We have therefore focused our effort on using this TAP purification protocol, using the pBBRDEST-42 plasmid as it encodes both the V5 and 6xhis tags. This approach has been incorporated as a part of standard protocols in a high throughput system and a panel of 200 *R. palustris* genes are being processed to serve as the pilot group for this automated approach.

As a benchmark for developing and evaluating affinity-based methods for isolating molecular machines, we carried out a conventional biochemical isolation (sucrose density gradient centrifugation) of the *R. palustris* ribosome, followed by both "bottom-up" and "top-down" mass spectrometric analysis of the protein components of this large, abundant complex. We have identified 53 of the 54 predicted protein components of the ribosome using by the "bottom-up" method, and obtained accurate intact masses of 42 ribosomal proteins using the "top-down" approach. Combining results from these two approaches provided information on post-translational modification of the ribosomal proteins, including N-terminal methionine truncation, methylation, and acetylation.

8

Bioinformatics and Computing in the Genomics:GTL Center for Molecular and Cellular Systems - LIMS and Mass Spectrometric Analysis of Proteome Data

F. W. Larimer[1] (larimerfw@ornl.gov), G. A. Anderson[2], K. J. Auberry[2], G. R. Kiebel[2], E. S. Mendoza[2], D. D. Schmoyer[1], and M. B. Shah[1]

[1]Oak Ridge National Laboratory, Oak Ridge, TN and [2]Pacific Northwest National Laboratory, Richland, WA

Scientists at the Oak Ridge National Laboratory/Pacific Northwest National Laboratory (ORNL/PNNL) Genomics:GTL (GTL) Center for Molecular and Cellular Systems are generating large quantities of experimental and computational data. We have developed a prototype Laboratory Information Management System (LIMS) for data and sample tracking of laboratory operations and processes in the various laboratories of the Center. We have also developed a mass spectrometry data analysis system for automating the mass spectrometry data capture and storage, and computational proteomic analysis of this data.

A Laboratory Information Management System for the GTL Center for Molecular and Cellular Systems. The Laboratory Information Management System (LIMS) for the GTL Center for Molecular and Cellular Systems is a central data repository for all information related to production and analysis of GTL samples. It maintains a detailed pedigree for each GTL sample by capturing processing parameters, protocols, stocks, tests and analytical results for the complete life cycle of the sample. Project and study data are also maintained to define each sample in the context of the research tasks that it supports.

The LIMS system is implemented using the Nautilus™ software from Thermo Electron Corporation. This software provides a comprehensive yet extensible framework for a LIMS that can be customized to meet the requirements of the GTL project. Nautilus uses client/server architecture to access data maintained in a central Oracle database and presents an interface based on the Windows Explorer paradigm. The latest Nautilus release includes Web access and this will be added to GTL LIMS in the near future.

The LIMS is configured by establishing workflows that parallel the processing steps completed in the laboratory. For each process it is necessary to define the laboratory environment (stocks, storage locations, instruments, protocols), identify the items to track, the process parameters to collect, the tests that will be conducted, and the test results that will be reported. This information is then used to develop LIMS workflows that will ensure the collection of all critical data.

The initial GTL LIMS system configuration has been completed. This required customization of Nautilus to include additional GTL data items such as primers, genes, and vectors, and programmatic extensions to do GTL specific tasks such as copying files to the central file server and displaying files stored on the central file server. Program extensions also had to be developed to handle some of the processing steps for stocks stored in 96-well plates.

Future plans for the LIMS include additional reporting capabilities, integration with the mass spec data analysis pipeline, barcode implementation, and refinement of the process workflows.

PRISM Mass Spectrometry Proteomic Data Analysis System. The Proteomics Research Information Storage and Management (PRISM) System manages the very large amounts of data generated by the mass spectroscopy facility and automatically performs the automated analytical processing that converts it into information about proteins that were observed in biological samples. PRISM also collects and maintains information about the biological samples and the laboratory protocols and procedures that were used to prepare them.

PRISM is composed of distributed software components that operate cooperatively on a network of commercially available PC computer systems. It uses several relational databases to hold information and a set of autonomous programs that interact with these databases to perform much of the automated file handling and information processing. A large and readily expandable data file storage space is provided by a set of storage servers. The basic database software is a commercial product, but the database schemata and content and the autonomous programs have all been developed in-house to meet the unique and continually evolving requirements of the MS facility.

PRISM has been in continuous operation since March 2000, and has been continually upgraded. There have been four major upgrade cycles, and numerous minor ones, including the addition of new functionality and the expansion of capacity as new instruments are added to the facility. Most recently, PRISM has been upgraded

to maintain inter-system tracking information for GTL samples and the ability to maintain and process them in their as-delivered format (96-well plates).

PRISM manages data and research results for all of the mass spec based proteomics studies in our laboratory; this includes over 100 research campaigns or lines of investigation. This research has resulted in 15334 datasets from a number of different mass spectrometers. These datasets have required 41491 separate analysis operations to extract peptide and protein identifications. The total raw data volume managed by PRISM is in excess of 15 Tera bytes. The current rate of production results in approximately 800 datasets per month with significant increases expected in FY04.

Data Abstraction Layer (DAL). The DAL is middleware that will provide a level of abstraction for any data storage system in the proteomics pipeline (LIMS, Freezer Software, PRISM, etc.). It will provide a generic interface for building tools and applications that require access to the experimental data and analysis results. It will also allow the pipeline data to be extended without making changes in the manner in which an application already looks at the data. For example, it could be used to facilitate a query performed utilizing proteomic data originating from both PNNL and ORNL. The DAL will be used to provide an interface to the pipeline data as required by selected bioinformatics/analysis tools.

9

Advanced Computational Methodologies for Protein Mass Spectral Data Analysis

Gordon Anderson[1] (gordon@pnl.gov), Joshua Adkins[1], Andrei Borziak[2], Robert Day[2], Tema Fridman[2], Andrey Gorin[2], Frank Larimer[2], Chandra Narasimhan[2], Jane Razumovskaya[2], Heidi Sophia[1], David Tabb[2], Edward Uberbacher[2], Inna Vokler[2], and Li Wang[2]

[1]Pacific Northwest National Laboratory, Richland, WA and [2]Oak Ridge National Laboratory, Oak Ridge, TN

Completed analysis of a variety of genomes has lead to a revolution in the methods and approaches of what was traditionally protein biochemistry. Now, an analysis of a variety of protein functions can be undertaken on a genome wide level. Among some of the most interesting and complicated functions of proteins is their nature to form higher order functional complexes. Using a combination of protein pull-down techniques and combined capillary liquid chromatography/mass spectrometry (LC/MS) as a sensitive detector for proteins, new protein complexes are being identified as part of the Center for Molecular and Cellular Systems. Computational tools are being developed to assist in the interpretation of these data. For example, complications in these data arise from non-specific and transient protein interactions. Imperfect bioinformatic tools for peptide identifications that lead to protein identifications found in these complex pull-downs is also a problem. We are using the clustering program, OmniViz, as a tool for discovery of protein complexes in this combination of complicating protein identifications. This includes the ability to view various experiments in a virtual 1D dimension gel format to aid biologists in looking at the results and adjustable features that can be used to compare different ratios of sensitivity and specificity in the putative protein complexes. We are automating the process, leading to standardized approaches and reports for protein components of complexes.

Improved scoring algorithms for matching theoretical tandem mass spectra of peptides to observed spectra are being developed to replace existing scoring algorithms such as that used by SEQUEST. The likelihood of matches can be estimated by probabilistic analysis of fragment ions matches rather than computationally expensive cross-correlation. This greatly improves the speed and accuracy of the peptide scoring system. A computational system for peptide charge determination has also been developed with 98% accuracy using statistical and neural network methods. This allows a several-fold speedup in calculation time without loss of information.

Methods for *de novo* sequencing to construct sequence tags from MS/MS data have also been developed using a statistical combination of informational elements including the peaks in the neighborhood of expected B and Y ions. The approach utilizes all informational content of a given MS/MS experimental data set, including peak intensities, weak and noisy peaks, and unusual fragments. The 'Probability Profile Method' is capable of recognizing ion types with good accuracy, making the identification of peptides significantly more reliable. The method requires a training database of previously resolved spectra, which are used to determine "neighborhood patterns" for peak categories that correspond to ion types (N- or C-terminus ions, their dehydrated fragments, etc.). The established patterns are applied to assign probabilities for experimental spectra peaks to fit into these categories. Using this model, a significant portion of peaks in a raw experimental spectrum can be identified with a high confidence. PPM can be used in a number of ways: as a filter for peptide database lookup approach to determine peptides with post-translational modifications or peptide complexes, *de novo* approach and tag determination.

10

High-Throughput Cloning, Expression and Purification of *Rhodopseudomonas palustris* and *Shewanella oneidensis* Affinity Tagged Fusion Proteins for Protein Complex Isolation

Dale A. Pelletier[1*] (pelletierda@ornl.gov), Linda Foote[1], Brian S. Hooker[2], Peter Hoyt[1], Stephen J. Kennel[1], Vladimir Kery[2], Chiann-Tso Lin[2], Tse-Yuan Lu[1], Lye Meng Markillie[2], and Liang Shi[2]

[1]Oak Ridge National Laboratory, Oak Ridge, TN and [2]Pacific Northwest National Laboratory, Richland, WA

This poster will describe the approaches and progress in the joint Oak Ridge National Laboratory/Pacific Northwest National Laboratory (ORNL/PNNL) Center for Molecular and Cellular Systems pilot project on protein complexes. We have adopted the following process design for isolation of protein complexes: (1) construct adaptable plasmids for expression in multiple organisms, (2) adapt standard gene primer design, (3) PCR amplify target genes, (4) clone into donor vector/expression vectors, and (5) express in selected organisms for exogenous and endogenous complex isolation.

We have developed software that designs appropriate PCR primers flanking each gene such that any gene can be amplified from genomic DNA. The resulting PCR products can be directly recombined into entry vectors. We have used the Gateway® cloning system (Invitrogen) to produce entry clones that can be recombined into our modified broad host range expression vectors which contain ori genes compatible with replication in a variety of bacterial hosts.

We have performed PCR amplification from host genomic DNA in 96-well format and shown, using generic conditions, that 60-70% of the reactions yield the predicted size products. We have previously demonstrated automated PCR amplification, cleanup and gel analysis using liquid handling robots and are transitioning to high-throughput hardware. We have successfully PCR amplified approximately 80 *R. palustris* genes and cloned 40 into expression vectors. Twenty of these constructs have been electroporated into *R. palustris*. To date high-throughput electroporation has not been implemented but such 96-well systems are commercially available and will be tested. Plans for automation at this step include an automated colony picker and subsequent robot directed plasmid preps for QA and long term cataloging and storage. We have also successfully cloned over 30 *S. oneidensis* genes into expression vectors using the Gateway® system. Over 20 of these constructs have been successfully expressed in both *E. coli* and *S. oneidensis*.

Expression in *E. coli* and in hosts *R. palustris* and *S. oneidensis* has to date been evaluated primarily using manual processes. Cell samples from relatively large cultures are lysed and IMAC or TAP isolations are completed followed by verification of product by SDS-PAGE and/or Western blot. Tagged *S. oneidensis* proteins expressed in *E. coli* and *S. oneidensis* for exogenous bait experiments are then purified using single-step IMAC on a Qiagen Biorobot 3000 LS. Milligram quantities of up to 12 proteins in parallel have been purified using this automated system.

Sandia National Laboratories

Carbon Sequestration in *Synechococcus*

From Molecular Machines to Hierarchical Modeling

11

Modeling Cellular Response

Mark D. Rintoul (rintoul@sandia.gov), Steve Plimpton, Alex Slepoy, and Shawn Means

Sandia National Laboratories, Albuquerque, NM

While much of the fundamental research on prokaryotes is focused on specific molecular mechanisms within the cell, the aggregate cellular response is also important to practical problems of interest. In this poster, we present results for two computational models of cellular response that take spatial effects into consideration. The first model is a discrete particle code where particles diffuse and interact via Monte Carlo rules so that species concentrations track chemical rate equations. This type of model is relevant to cases where there are a small number of interacting particles, and the spatial and temporal fluctuations in particle number can play a significant role in affecting cellular response. Results with this code are shown for a simulation of the carbon sequestration process in *Synechococcus WH810*. The second model utilizes continuum modeling focusing on carbon concentrations inside and outside of the cell, in an effort to understand carbon transport by *Synechococcu*s within a fluid-dynamic marine environment. It is based on solving partial differential equations on a realistic geometry using finite element methods.

12

The *Synechococcus* Encyclopedia

Nagiza F. Samatova[1], **Al Geist**[1] (gst@ornl.gov), Praveen Chandramohan[1], Ramya Krishnamurthy[1], Gong-Xin Yu[1], and Grant Heffelfinger[2]

[1]Oak Ridge National Laboratory, Oak Ridge, TN and [2]Sandia National Laboratories, Albuquerque, NM

Synechococcus sp. are abundant marine cyanobacteria known to be important to global carbon fixation. Although the genome sequencing of *Synechococcus sp*. is complete by the DOE JGI[1], the actual biochemical mechanisms of carbon fixation and their genomic basis are poorly understood. This topic is under both experimental and computational investigation by several projects including Dr. Brian Palenik's DOE MCP project, SNL/ORNL GTL Center[2] and others. These projects have been generating heterogeneous data (e.g., sequence, structure, biochemical, physiological and genetic data) distributed across various institutions. Integrative analysis of these data will yield major insights into the carbon sequestration behavior of

Synechococcus sp. However, such analysis is largely hampered by the lack of a knowledgebase system that enables an efficient access, management, curation, and computation with these data as well as comparative analysis with other microbial genomes. To fulfill these requirements, the *Synechococcus Encyclopedia* is being created as part of the SNL/ORNL GTL Center.

The completed sequencing of *Synechococcus sp.* has allowed having a reference axis upon which any type of annotation can be layered. Not only can genomic features such as genes and repeats be placed upon such a reference, but it is also possible to map a variety of other data such as operons and regulons, mutations, phenotypes, gene expressions (e.g., microarray, phage display, mass spec, 2-hybrid), pathway models, protein interactions, and structures. A major benefit of such feature mapping is that each of these annotations can be cross-referenced to each other. The *Synechococcus* Encyclopedia takes advantage of this fact to allow users to view and track a variety of biological information associated with the genome and to enable complex queries across multiple data types.

In order to make the exploration and in-depth analysis of genome information easier, one needs appropriate ways to browse and query the corresponding data. The World Wide Web interface of the *Synechococcus* Encyclopedia was built up with these specifications in mind. It offers a number of ways to retrieve information about a genome. For instance, an advanced search capability is available to combine several search criteria and retrieve detailed information about any intricate features. Moreover, the search can be restricted to a genome region of interest, molecular function, biological process, or cellular component.

To ease data retrieval, all output reports are presented in tabular format and maybe conveniently downloaded as tab-delimited text. If desired, the tables can be easily customized, e.g. adding or removing features, or changing the sort order according to several data fields. Data are accessible through a variety of interactive graphical viewers. Furthermore, the retrieved data entries can be further explored by launching complex analysis tools or linking to other data collections such as Swiss-Prot, Pfam, InterPro, PDB, etc.

The *Synechococcus* Encyclopedia comprises information at various levels: genome sequence, structure, regulation, protein interactions, systems biology. For example, at the genome sequence and annotation level, it includes the protein- and RNA-coding genes, Pfam domains, Blocks motifs, InterPro signatures, and COG- and KEGG-based functional assignments. At the structure level, it presents secondary and tertiary structural models predicted by PROSPECT[3] and other tools, SCOP-based functional assignments, and FSSP profiles for homologous protein sequences. At the regulation level, it integrates data about promoters, transcription factors, and pathway models as well as microarray data.

The *Synechococcus* Encyclopedia is accessible at http://www.genomes-to-life.org/syn_wh. The generic data model and data integration, search and retrieval engine that it is based on, makes it possible to set up similar knowledgebases for other bacterial species. New functionalities for multi-genome integration and comparative analysis of genomes are being developed to facilitate better understanding of genomic organization and biological function. Moreover, other features such as annotation and curation services, data provenance, and security will be added in the near future.

References

1. http://genome.ornl.gov/microbial/syn_wh

2. Grant Heffelfinger (gsheffe@sandia.gov) and Al Geist (gst@ornl.gov); http://www.genomes2life.org

3. Ying Xu (xyn@ornl.gov) and Dong Xu (xudong@missouri.edu); http://compbio.ornl.gov/structure/prospect/

13

Carbon Sequestration in *Synechococcus*: A Computational Biology Approach to Relate the Genome to Ecosystem Response

Grant S. Heffelfinger (gsheffe@sandia.gov)

Sandia National Laboratories, Albuquerque, NM

This talk will provide an update on the progress to date of the Genomics:GTL (GTL) project led by Sandia National Laboratories: "Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling." This effort is focused on developing, prototyping, and applying new computational tools and methods to elucidate 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 project includes both an experimental investigation as well as significant computational efforts to develop and prototype new computational biology tools. Several elements of this effort will be discussed including the development of new methods for high-throughput discovery and characterization of protein-protein complexes and novel capabilities for inference of regulatory pathways in microbial genomes across multiple sources of information. Our progress developing new computational systems-biology methods for understanding the carbon fixation behavior of *Synechococcus* at different levels of resolution from the cellular level to ecosystem will also be discussed. More information about our project and partners can be found at www.genomes-to-life.org.

14

Improving Microarray Analysis with Hyperspectral Imaging, Experimental Design, and Multivariate Data Analysis

David M. Haaland[1] (dmhaala@sandia.gov), Jerilyn A. Timlin[1], Michael B. Sinclair[1], Mark H. Van Benthem[1], Michael R. Keenan[1], Edward V. Thomas[1], M. Juanita Martinez[2], Margaret Werner-Washburne[2], Brian Palenik[3], and Ian Paulsen[4]

[1]Sandia National Laboratories, Albuquerque, NM; [2]University of New Mexico, Albuquerque, NM; [3]Scripps Institution of Oceanography, La Jolla, CA; and [4]The Institute for Genomic Research, Rockville, MD

At Sandia National Laboratories, we are combining hyperspectral microarray scanning, efficient experimental designs, and a variety of new multivariate analysis approaches to improve the quality of data and the information content obtained from microarray experiments. Our approach is designed to impact the Sandia-led GTL team's investigation of *Synechococcus* for carbon sequestration. Current commercial microarray scanners use univariate methods to quantify a small number of dyes on printed microarray slides. We have developed a new hyperspectral microarray scanning system that offers higher throughput for each microarray slide by allowing the quantitation of a large number of dyes on each slide. The new scanner has demonstrated improved accuracy, precision, and reliability in quantifying dyes on microarrays and yields a higher dynamic range than possible with current commercial scanners. We will present the design of the new scanner, which collects the entire fluorescence spectrum from each pixel of the scanned microarray, and the use of multivariate curve resolution (MCR) algorithms to obtain pure emission spectra and corresponding concentration maps from the hyperspectral image data. The new scanner has allowed us to detect contaminating autofluorescence that emits at the same wavelengths as the reporter fluorophores on microarray slides. With the new scanner, we are able to generate relative concentration maps of the background, impurity, and fluorescent labels at each pixel of the image. Since the MCR generated concentration maps of the fluorescent labels are unaffected by the presence of background and impurity emissions, the accuracy and useful dynamic range of the gene expression data are both greatly improved. We will also demonstrate that the new scanner helps us understand a variety of artifacts that have been observed with microarrays scanned using two-color scanners. Artifacts include high background intensities, "black holes," dye separation, the presence of unincorporated dye, and contaminants that have led to the practice of intensity-dependent normalizations.

We will describe statistically designed microarray experimental approaches that we have used to identify and eliminate experimental error sources in the microarray technology. These statistically designed experiments have led to dramatic improvements in the quality and reproducibility of yeast microarray experiments. The lessons learned from yeast arrays will be applied directly to our GTL investigations of *Synechococcus* microarrays. In addition, new approaches with multivariate algorithms that incorporate error covariance of the arrays into the multivariate analysis of microarrays will be presented along with methods to evaluate the relative performance of various gene selection, classification, and multivariate fitting algorithms.

Evaluation of hybridization experiments with initial 250 gene *Synechococcus* microarrays will be presented along with graphical methods designed to facilitate understanding of the quality and repeatability of the *Synechococcus* microarray data. Work has recently begun on the whole genome *Synechococcus* microarray experi-

ments. cDNA arrays of 2496 genes from *Synechococcus* have been printed on glass slides. We will present the unique design of the whole genome arrays which includes six replicates for each gene each printed with a different pin to capture the true within array repeatability of gene expression. The arrays also include multiple positive controls, negative controls, blanks, and solvent spots in each block of the whole genome microarray. In addition, Arabidopsis promoter 70mers were printed on the four corners of each block to assist in positioning and reading the array. If available at the time of the workshop, gene expression results from the whole genome *Synechococcus* microarrays will be presented.

Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-ACO4-94AL85000. This work was funded in part by the US Dep't of Energy's Genomics:GTL program (www.doegenomestolife.org) under project, "Carbon Sequestration in *Synechococcus Sp*.: From Molecular Machines to Hierarchical Modeling," (http://www.genomes2life.org/). This project also supported in part by a grant from the W. M. Keck Foundation.

15

Multi-Resolution Functional Characterization of *Synechococcus* WH8102

Nagiza F. Samatova*[1] (samatovan@ornl.gov), Andrea Belgrano[9], Praveen Chandramohan[1], Pan Chongle[1], Paul S. Crozier[2], Al Geist[1], Damian Gessler[9], Andrey Gorin[1], Jean-Loup Faulon[2], Hashim M. Al-Hashimi[7], Eric Jakobbson[4], Elebeoba May[2], Anthony Martino[2], Shawn Means[2], Rajesh Munavalli[1], George Ostrouchov[1], Brian Palenik[5], Byung-Hoon Park[1], Susan Rempe[2], Mark D. Rintoul[2], Diana Roe[2], Peter Steadman[9], Charlie E. M. Strauss[3], Jerilyn Timlin[2], Gong-Xin Yu[1], Maggie Werner-Washburne[10], Dong Xu[8], Ying Xu[6], and **Grant Heffelfinger**[2] (gsheffe@sandia.gov)

*Presenting author

[1]Oak Ridge National Laboratory, Oak Ridge, TN; [2]Sandia National Laboratories, Albuquerque, NM and Livermore, CA; [3]Los Alamos National Laboratory, Los Alamos, NM; [4]University of Illinois, Urbana Champaign, IL; [5]Scripps Institution of Oceanography, La Jolla, CA; [6]University of Georgia, Atlanta, GA; [7]University of Michigan, Ann Arbor, MI; [8]University of Missouri, Columbia, MO; [9]National Center for Genome Resources, Santa Fe, NM; and [10]University of New Mexico, Albuquerque, NM

Although sequencing of multi-megabase regions of DNA has become quite routine, it remains a big challenge to characterize all the segments of DNA sequence with various biological roles such as encoding proteins and RNA or controlling when and where those molecules are expressed. The primary difficulty is that function exists at many hierarchical levels of description, it has temporal and spatial connections that are difficult to manage, and functional descriptions do not correspond to well-defined physical models like biological structures defined by Cartesian coordinates for atoms. Results from the completed prokaryotic genome sequences show that almost half of the predicted coding regions identified are of unknown biological function. Specifically, the completed sequencing of *Synechococcus* WH8102 by JGI and multi-institutional annotation effort[1] has resulted in 1196 (out of 2522) ORFs that are conserved hypothetical or hypothetical.

The goal of this work is to develop a suite of computational tools for systematic multi-resolution functional characterization of microbial genomes and utilize them

to elucidate the biochemical mechanisms of the carbon sequestration of *Synechococcus sp.* as part of the SNL/ORNL GTL Center[2]. Functional characterization is conducted on many levels and with different questions in mind – ranging from the reconstruction of genome-wide protein-protein interaction networks 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. Here we outline some of them and provide pointers to more information.

While analysis of a single genome provides tremendous biological insights on any given organism, comparative analysis of multiple genomes can provide substantially more information on the physiology and evolution of microbial species. Comparative studies expand our ability to better assign putative function to predicted coding sequences and our ability to discover novel genes and biochemical pathways. The KeyGeneMiner[3] aims to identify "key" genes that are responsible for a given biochemical process of interest. When applied to the oxygenic photosynthetic process, it has discovered 126 genome features. Many of them have been reported in literature as either photosynthesis-related or photosynthesis-specific (occurring only in photosynthetic genomes). Likewise, the construction of comprehensive phylogenetic profiles for all transport proteins in the bacterial genomes[4] allows us to pick up regulators of transport and help annotate some genes for which there is still no, or weak, annotation. The approach is not limited to transport proteins; it can be done with any set of probe sequences representing an interesting functional grouping.

After genes are assigned to putative biochemical processes or putative functional links are established between genes, there still remains a significant challenge to understand how these biomolecules interact to form pathways for metabolic conversion from one substance to another and how genes form networks to regulate the timing and location events within the cell. In spite of some promising work using Boolean and Bayesian networks, all these approaches are challenged with too many parameters (compared to the number of data points) to adequately constrain the problem thus resulting in too many plausible competing gene models. The complexity of this problem is begging for novel methods that could integrate information from appropriate databases (e.g., gene expression, protein-protein interaction data, operon and regulon structure, transcription factors) in order to constrain the set of plausible solutions as well as to use this information for designing targeted experiments for study of specific network modules. Our progress towards this goal includes the development of new and effective protocols for systematic characterization of regulatory pathways and a preliminary version of a computational pipeline for interpretation of multiple types of biological data for biological pathway inference[5]. We have also prototyped these methods on *Synechococcus* WH8102 to make several predictions, including a signaling/regulatory network for the phosphorus assimilation pathway. We have developed an environment[6] to aid biologists in the analysis of proposed networks via visual browsing of the annotation information and original data associated with different elements of these networks. Using these tools and large visualization corridor with 48 high-resolution screens we have been able to begin assigning biological processes to groups of genes which were aggregated together by similar expression, as measured by microarrays, and then placed into a Boolean network based on discrete (Boolean) expression levels.

Structural characterization of protein machines provides additional valuable insights about the mechanism and details of their function. In spite of a long history behind the development of computational methods for structural characterization of protein machines, many challenges still remain to be addressed. Specifically, we are focusing on the methods for understanding how biological molecules interact physi-

cally to transfer signals, including protein-protein interactions as well as protein-DNA and protein-RNA interactions[7]. At the initial stage we apply structure prediction methods to determine protein fold families with our ROSETTA[8] and PROSPECT[9] programs and use inferred structural similarities to create hypotheses about their interacting partners. The computational pipeline merges several bioinformatics and modeling tools including algorithms for protein domain division, secondary structure prediction, fragment library assembly, and structure comparison[7]. Application of this pipeline has resulted in genome-scale structural models for *Synechococcus* genes[10].

Protein interactions with other molecules play a central role in determining the functions of proteins in biological systems. Protein-protein, protein-DNA, protein-RNA, and enzymes-substrate interactions are a subset of these interactions that is of key importance in metabolic, signaling and regulatory pathways. Bacterial chemotaxis, osmoregulation, carbon fixation and nitrogen metabolism are just a few examples of the many complex processes dominated by such molecular recognition. Identification of interacting proteins is an important prerequisite step in understanding its physiological function. We develop several complimentary approaches to this problem including statistically significant protein profiles based[3], signature kernel SVMs based[11], and genomic context based[5]. Utilization of these approaches produced a genomes-scale protein interaction map for *Synechococcus* WH8102.

Knowing interacting partners is just the first step towards elucidating the order and control principles of molecular recognition. One of the most remarkable properties of protein interactions is high specificity. Even presumably specific binding sites may bind a range of ligands with different compositions and shapes. For example, the Ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco) can catalyze two separate reactions, carboxilation and oxygenation reactions depending on whether CO_2 or O_2 binds in its active site, respectively. There must be then *functionally important residues* that enable different proteins to recognize their unique interacting partners. Identification of these functionally important residues (e.g. docking interfaces, catalytic centers, substrate and cofactor binding sites and hinge-motion controlling loops) is essential for functional characterization of protein machines. We explore several complimentary approaches to identify protein docking interfaces based on quantification of correlated mutations[3], Boosting driven separation into "predictable" feature subspaces[3], and statistically significant separation of likely *n*-mers[6]. Likewise, our Surface Patch Ranking method[3] has been utilized to identify clusters of residues important for CO_2/O_2 specificity in Rubisco. Finally, our tools for full atom modeling of protein interactions (LAMMPS, PDOCK)[10] are being used to assess Rubisco catalytic specificity.

All these tools and experiments generate heterogeneous data (e.g., sequence, structure, biochemical, physiological and genetic data). Integrative analysis of these data will yield major insights into the carbon sequestration behavior of *Synechococcus sp*. For this to occur, low level molecular data and processes need to be connected to macro-ecological models. We are doing this in a Hierarchical Simulation Platform[12] that uses recently published models connecting metabolic rate to population growth and trophic level energy flux. Additionally, we have built a web portal for *Synechococcus* Encyclopedia[13] that enables an efficient access, management, curation, and computation with these data as well as comparative analysis with other microbial genomes.

References

1. http://genome.ornl.gov/microbial/syn_wh

2. Grant Heffelfinger (gsheffe@sandia.gov) and Al Geist (gst@ornl.gov); http://www.genomes2life.org

3. Nagiza Samatova (samatovan@ornl.gov)

4. Eric Jakobsson (jake@ncsa.uiuc.edu)

5. Ying Xu (xyn@ornl.gov), Brian Palenik (bpalenik@ucsd.edu), and Dave Haaland (dmhaala@sandia.gov); see "Microarray Analysis with Hyperspectral Imaging, Experimental Design, and Multivariate Data Analysis" and "Methods for Ellucidating Synechococcus Regulatory Pathways" posters

6. George S Davidson (gsdavid@sandia.gov)

7. Andrey Gorin (agor@ornl.gov); see also "Bioinformatics Methods for Mass Spect Analysis" poster

8. Charlie Strauss (cems@lanl.gov)

9. Ying Xu (xyn@ornl.gov) and Dong Xu (xudong@missouri.edu); http://compbio.ornl.gov/structure/prospect/

10. Ying Xu (xyn@ornl.gov); http://compbio.ornl.gov/PROSPECT/syn/

11. Daniel M Rintoul (mdrinto@sandia.gov) and Antony Martino (martino@sandia.gov); see also "Modeling Cellular Response" poster

12. Damian Gessler (ddg@ncgr.org), Andrea Belgrano (ab@ncgr.org), and Peter Steadman (ps@ncgr.org).

13. Al Geist (gst@ornl.gov) and Nagiza Samatova (samatovan@ornl.gov); http://www.genomes-to-life.org/syn_wh; see also "The *Synechococcus* Encyclopedia" poster.

16

Computational Inference of Regulatory Networks in *Synechococcus sp* WH8102

Zhengchang Su[1], Phuongan Dam[1], Hanchuan Peng[1], **Ying Xu**[1] (xyn@bmb.uga.edu), Xin Chen[2], Tao Jiang[2], Dong Xu[3], Xuefeng Wan[3], and Brian Palenik[4]

[1]University of Georgia, Athens, GA and Oak Ridge National Laboratory, Oak Ridge, TN; [2]University of California, Riverside, CA; [3]University of Missouri, Columbia, MO; and [4]University of California, San Diego, CA

In living systems, control of biological function occurs at the cellular and molecular levels. These controls are implemented by the regulation of activities and concentrations of species taking part in biochemical reactions. The complex machinery for transmitting and implementing the regulatory signals is made of a network of interacting proteins, called *regulatory networks*. Characterization of these regulatory networks or pathways is essential to our understanding of biological functions at both molecular and cellular levels.

We have been developing a prototype system for computational inference of regulatory and signaling pathways for the genome of *Synechococcus sp.* WH8102. Currently, the prototype system consists of the following components: (a) prediction of gene

functions, (b) prediction of terminators of operon structures, (c) genome-scale prediction of operon structures, (d) genome-scale prediction of regulatory binding sites, (e) mapping of orthologous genes and biological pathways across related microbial genomes, (f) prediction of protein-protein interactions, (g) mapping biological pathways across related genomes, and (h) inference of pathway models through fusing the information collected in steps (a) through (g).

a. **Computational prediction of gene functions.** We have previously developed a computational pipeline for inference of protein structures and functions at genome scale (Shah, et al. 2003 and Xu, et al. 2003). The pipeline consists of both sequence-based homology detection programs like psi-BLAST, and structure-based homology detection program PROSPECT (Xu, et al. 2000). We found that using structure-based approach in addition to psi-BLAST, we can detect additional 10-20% of remote homologs for genes in a microbial genome. This pipeline can be accessed at http://compbio.ornl.gov/PROSPECT/PROSPECT-Pipeline/cgi-bin/proteinpipeline_form.cgi. We have applied this pipeline to all the orfs of *Synechococcus sp* WH8102, and assigned close to 80% of its genes to some level of functions. All the results can be found at http://compbio.ornl.gov/PROSPECT/syn/.

b. **Prediction of terminators of operons.** We are in the early phase of developing a computational capability for prediction of terminators in WH8102. Our initial focus has been on *rho*-independent terminators (RIT). We are carrying out a comparative genome analysis to compare the RITs of the otthologous genes in different genomes to identify possible conserved patterns. We are also developing and implementing a novel algorithm based on MST clustering approaches to use common features of RITs to predict new RITs. We will apply more sophisticated energy functions than the one used in TransTerm and RNAMotif.

c. **Prediction of operons.** We have been working on a comparative genomics approach for predicting operons in *Synechococcus sp*. WH8102 that combines many known characteristics of an operon structure concerning the functions, intergenic distances and transcriptional directions of genes, promoters, terminators, etc. in a unified likelihood framework (Chen, et al. 2003). The data and results are available to the public at http://www.cs.ucr.edu/~xinchen/operons.htm. We have used the predicted operons, as one piece of information, in our inference of regulatory pathways in *Synechococcus sp* WH8102.

d. **Prediction of regulatory binding sites.** We have previously developed a computer program CUBIC for identification of consensus sequence motifs as possible regulatory binding sites (Olman et al. 2003). CUBIC solves the binding site identification problem as a problem of identifying data clusters from a noisy background. We have applied CUBIC for binding site predictions at genome scale, through identifying orthologous genes of WH8102 in other related genomes and application of CUBIC to the upstream regions of each set of orthologous genes. We expect that the genome-scale binding site prediction results will be publicly available within weeks.

e. **Mapping of orthologous genes across related genomes.** The identification of orthologous genes is a fundamental problem in comparative genomics and evolution, and is very challenging especially on a genome-scale. We have been working on anew approach for assigning orthologs between different (but related) genomes based on homology search and genome rearrangement. The preliminary experimental results on simulated and real data demonstrate that

the approach is very promising (it is competitive to the existing methods), although more needs to be done (Chen, et al. 2004).

f. Prediction of protein-protein interactions. We have implemented a computer software for predicting protein-protein interactions, employing a number of popular prediction strategies, including mapping against protein-protein interaction maps derived from experiments (like two hybrid), application of phylogenetic profile analysis (Pellegrini et al. 1999) and gene fusion method (Marcotte et al. 1999). We have made a genome-scale prediction of protein-protein interactions for *Synechococccus sp* genes.

g. Mapping of biological pathways across related genomes. We have recently developed a computational method for mapping of biological pathways across related microbial genomes. The core component of the algorithm/program is orthologous gene mapping under the constraints of (a) operon structures, (b) regulon structures (defined in terms of operons with common regulatory binding sites), and (c) co-expressions of genes. We have implemented this algorithm as a computer program P-MAP, and apply this program to assign all known pathways in *E coli*. (partial or complete) to *Synechococcus sp*. WH8102. The mapping results will soon be posted at our *Synechococcus* Knowledge Dabase at http://csbl.bmb.uga.edu/~peng/home.html.

h. Pathway inference through information. We have developed a computational protocol for inference of regulatory and signaling pathways through fusing the information collected in the steps (a) through (g) and a simple merging-voting scheme to put the predicted complexes, protein-DNA interactions and protein-protein interactions. We have applied this capability to the inference to the prediction of a number of regulatory pathways, including phosphorus assimilation pathway [ref], nitrogen and carbon assimilation pathways [unpublished results] of *Synechococcus sp* WH8102. We are currently exploring a number of formalisms for piecing together predicted gene associations into pathway models, including Biochemical Systems Theory (BST) (Savageau 1976).

Experimental validations of predictions are being carried out using microarray analyses (see Haaland et al poster) of wild type WH8102 and knockout mutants under selected growth conditions.

ACKNOWLEDGEMENTS. This work is funded in part by the US Department of Energy's Genomics:GTL (www.doegenomestolife.org) under project "Carbon Sequestration in *Synechococcus sp*: From Molecular Machines to Hierarchcal Modeling" (www.genomes-to-life.org).

References

1. X. Chen, Z. Su, P. Dam, B. Palenik, Y. Xu, and T. Jiang. Operon prediction by comparative genomics: an application to the Synechochoccus sp. WH8102 genome. 2003, submitted to *Nuc. Acids Res*. (in revision).

2. X. Chen, J. Zheng, Z. Fu, P. Nan, Y. Zhong, S. Lonardi, and T. Jiang. Assignment of orthologous genes via genome rearrangement. 2004, submitted to ISMB'2004.

3. E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice et.al, Detecting protein function and protein-protein interactions from genome sequences, *Science*, **285**:751-753, 1999.

4. V. Olman, D. Xu and Ying Xu, "Identification of Regulatory Binding-sites using Minimum Spanning Trees", Proceedings of the 7th Pacific Symposium on Biocomputing (PSB), pp 327-338, 2003.

5.

M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg et.al, Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, *Proc Natl Acad Sci USA*, **96**:4285-4288, 1999.

6. M. Shah, S. Passovets, D. Kim, K. Ellrott, L. Wang, I. Vokler, P. Locascio, D. Xu, Ying Xu, A Computational Pipeline for Protein Structure Prediction and Analysis at Genome Scale, *Proceedings of IEEE Conference on Bioinformatics and Biotechnology*, 3-10, IEEE/CS Press, 2003 (An expanded journal version is published in *Bioinformatics*, **19**(15):1985-1996, 2003).

7. M. A. Savageau. "Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology," Addison-Wesley, Reading, Mass (1976).

8. Z. Su, A. Dam, X. Chen, V Olman, T. Jiang, B. Palenik, and Ying Xu, Computational Inference of Regulatory Pathways in Microbes: an application to the construction of phosphorus assimilation pathways in Synechococcus WH8102, Proceedings of 14th *International Conference on Genome Informatics* pp:3-13,Universal Academy Publishing, 2003.

9. D. Xu, P. Dam, D. Kim, M. Shah, E. Uberbacher, and Ying Xu, Characterization of Protein Structrue and Function at Genome-scale using a Computational Prediction Pipeline, accepted to appear in *Genetic Engineering: Principles and Methods*, Vol **32**, Jane Setlow (Ed.), Plenum Press, 2003.

10. Y. Xu and D. Xu. Protein threading using PROSPECT: Design and evaluation. *Proteins: Structure, Function, and Genetics*. **40**:343-354. 2000.

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

17

Analysis of Predominant Genome Sequences and Gene Expression During *In Situ* Uranium Bioremediation and Harvesting Electricity from Waste Organic Matter

Stacy Ciufo[1]*, Dawn Holmes[1], Zhenya Shelbolina[1], Barbara Methé[2], Kelly Nevin[1], and **Derek Lovley**[1] (dlovley@microbio.umass.edu)

*Presenting author

[1]Department of Microbiology, University of Massachusetts, Amherst, MA and [2]The Institute for Genomic Research, Rockville, MD

Field studies have demonstrated that stimulating dissimilatory metal reduction in uranium-contaminated subsurface environments is an effective, simple, and inexpensive method for removing uranium from contaminated groundwater. Molecular analyses, which avoid any culture bias, have demonstrated that *Geobacter* species are the predominant microorganisms in a variety of subsurface environments in which dissimilatory metal reduction is an important process. For example, during *in situ* bioremediation of a uranium-contaminated site in Rifle, Colorado *Geobacter* species accounted for as much as 80% of the microbial community in the groundwater during uranium bioremediation. In a similar manner, it has been demonstrated that *Geobacter* species are the predominant microorganisms on electrodes used to harvest electricity from waste organic matter. In order to determine whether models of *Geobacter* physiology derived from pure culture studies are applicable to as-yet-uncultured *Geobacter*s living in uranium-contaminated subsurface environments or on the surface of electrodes, it is necessary to determine the relative similarities of the genome sequence and gene expression patterns of as-yet-uncultured *Geobacter*s and the pure cultures.

One strategy to evaluate the genetic potential of the *Geobacter*s that predominate in the environments of interest is to sequence genomic DNA directly extracted from the environment. Genomic DNA was extracted from the sediments of a uranium-contaminated aquifer, located in Rifle, Colorado, in which the activity of *Geobacter*s had been stimulated with the addition of acetate to promote precipitation of uranium. The DNA was cloned in bacterial artificial chromosomes (BACs) with an average insert size of ca. 40 kbp. Large scale sequencing of the BAC inserts resulted in the recovery of 4.2 mbp of environmental genomic DNA sequence. The sequence data was assembled with BACPACK, an algorithm we specifically developed for this purpose. A contig of over 580 kbp of sequence was assembled as were

several other contigs of 25-475 kbp. Analysis for highly conserved *Geobacter* genes indicated that these contigs were from a *Geobacter* species. The uncultured *Geobacter* had a 16S rRNA gene sequence identical to a sequence that predominated in the groundwater during uranium bioremediation. The environmental genome sequence was similar to that of pure *Geobacter* species in that it had a lower percentage of putative proteins predicted to be localized in the cytoplasm and more proteins targeted to the inner membrane, periplasm, and outer membrane than has been found in non-*Geobacteraceae*. The uncultured *Geobacter* species had a high percentage of *Geobacter* signature genes and gene arrangements similar to those found in cultured species. However, there were also genes of unknown function in the uncultured *Geobacter* that have not been identified in the genomes of any pure cultures. These results suggest that although the uncultured *Geobacter* species involved in uranium bioremediation are clearly not identical to the pure cultures that are being intensively studied, there are many genomic similarities and thus models of pure cultures may have applicability to *Geobacter*-dominated subsurface environments.

Another strategy to determine the genome sequences of the *Geobacter*s predominating in environments of interest is to adapt culture conditions to permit culturing of these organisms. A medium in which the clay-fraction of subsurface sediments was the source of Fe(III) oxide was developed. With this medium a *Geobacter* species with a 16S rRNA gene sequence identical to one of the sequences that predominated during uranium bioremediation was isolated. Cultivation of this organism required the addition of groundwater from the site to the medium. Sufficient quantities of this organism have now been cultured so that its genome can be sequenced. Of further interest is the finding that one of the large BAC contigs from the uranium bioremediation site has the same 16S rRNA gene sequence. Thus, it will be possible to compare results from direct sequencing of environmental genomic DNA with the strategy of isolation in culture followed by genome sequencing.

One test of the environmental applicability of the current physiological models of *Geobacter* metabolism is to determine whether *Geobacters* in environments of interest have patterns of gene expression that are similar to those in pure cultures. Therefore, methods for effectively extracting mRNA from aquifer sediments and the surface of energy-harvesting electrodes were developed. Initial studies on the metabolic state of *Geobacter* species in aquifer sediments demonstrated that the natural populations of *Geobacters* were highly expressing genes for nitrogen fixation, suggesting that they were limited for fixed nitrogen. Addition of 100 μM ammonium to the sediment repressed expression of the nitrogen fixation genes. These results demonstrated that it is possible to evaluate the *in situ* metabolic state of *Geobacters* in subsurface environments. The next step will be to evaluate the expression of a larger suite of genes involved in nutrient uptake and stress response with microarrays.

Evaluation of gene expression in *Geobacter sulfurreducens* growing on the surface of energy-harvesting electrodes suggested that environmental analysis with whole-genome DNA microarrays is feasible. A microarray was used to compare mRNA levels of *G. sulfurreducens* growing on electrodes with mRNA levels of planktonic cells. Up-regulation of several genes was significant on the electrode. For example, mRNA levels for several outer-membrane cytochromes were 40-80 fold higher in cells growing on the electrodes. This suggests that these cytochromes play an important role in electron transfer to the electrode surfaces. There was also an upregulation of genes annotated as encoding for heavy-metal efflux proteins. This may reflect the presence of heavy metal contaminants in the electrode material. Many genes of unknown function were also down-regulated. The most prominently down-regulated genes were related to oxygen respiration and/or oxygen toxicity. These included a cytochrome oxidase as well as thioredoxin peroxidase and

superoxide dismutase. This is indicative of an important change in the respiratory pathway. These results demonstrate that electron transfer to electrodes is associated with significant shifts in gene expression and provide the first insights into the mechanisms for this novel form of respiration.

In summary, these initial results demonstrate that it will be possible not only to determine the genetic potential of the *Geobacter* species actually involved in subsurface bioremediation or in harvesting electricity from waste organic matter, but also to broadly assess their metabolic state. This will significantly improve the development of *in silico* models for predicting the metabolic responses of *Geobacter* species under different environmental conditions and provide information on how to most effectively optimize these applications of *Geobacter*.

18

Functional Analysis of Genes Involved in Electron Transport to Metals in *Geobacter sulfurreducens*

Maddalena Coppi[1]*, Eman Afkar[1], Tunde Mester[1], Daniel Bond[1], Laurie DiDonato[1], Byoung-Chan Kim[1], Richard Glaven[1], Ching Leang[1], Winston Lin[1], Jessica Butler[1], Teena Mehta[1], Susan Childers[1], Barbara Methé[2], Kelly Nevin[1], and **Derek Lovley**[1] (dlovley@microbio.umass.edu)

*Presenting author

[1]Department of Microbiology, University of Massachusetts, Amherst, MA and [2]The Institute for Genomic Research, Rockville, MD

As noted in a companion abstract, ecological studies have demonstrated that *Geobacter* species are the predominant microorganisms in a variety of subsurface environments in which dissimilatory metal reduction is an important process, including during *in situ* uranium bioremediation. Therefore, in order to effectively model *in situ* bioremediation of uranium and develop strategies for improving this process it is necessary to understand the factors controlling the growth and activity of *Geobacter* species. Most important in this regard is information on electron transfer not only to U(VI), but also to Fe(III), because most of the energy supporting the growth of *Geobacter* species in uranium-contaminated subsurface environments is derived from electron transfer to Fe(III).

Functional analysis of electron transfer to metals in *Geobacter sulfurreducens* has initially focused on the *c*-type cytochromes which are abundant in the genome, as well as outer-membrane proteins of previously unknown function. For example, analysis of the outer-membrane proteins of *G. sulfurreducens* with MALDI-TOF mass spectrometry revealed that the most abundant protein, designated OmpA, had a predicted amino acid sequence without any significant homology with previously described genes. OmpA is predicted to have a hydrophobic leader sequence, consistent with export to the outer membrane, and a β-barrel structure. No heme *c* or metal binding motifs were detected. When the gene was deleted with the single gene replacement method, the *ompA*-deficient mutant grew the same as wild type with fumarate as the electron acceptor, but it could not grow with Fe(III) or Mn(IV) oxides as the electron acceptor. Although the total heme *c* content in the mutant and the wild type were comparable, the mutant had only ca. 50% of the heme *c* content in the outer membrane as the wild type. There was a corresponding substantial increase in the total heme *c* content in the cytoplasmic membrane and

soluble fraction of the *ompA* mutant. These results suggest that OmpA plays an important role in localizing *c*-type cytochromes in the outer membrane of *G. sulfurreducens* via a novel mechanism not previously described in any microorganism.

A mutation in a novel secretory system in *G. sulfurreducens* specifically eliminated its ability to reduce Fe(III) oxides, but not soluble electron acceptors, including chelated Fe(III). Comparison of the proteins in the periplasm of this mutant with wild type cells indicated that several proteins were accumulating in the periplasm of the mutant. Analysis of peptide fragments of one of these proteins revealed a gene, designated *ompB*, which encodes for a 1303 amino acid protein, with 23 transmembrane amino acids and 1275 amino acids predicted to be exposed outside the cell. There are four putative metal-binding sites. This gene is found in the four *Geobacteraceae* genome sequences that are available, but not in any other organisms. The *ompB* mutant did not grow on Fe(III) oxide, but grew on soluble electron acceptors. These results suggest that OmpB plays an important role in cell-Fe(III) oxide contact or in sequestering Fe(III) from Fe(III) oxides, prior to Fe(III) reduction. This is a novel concept for dissimilatory Fe(III) oxide reduction.

Our previous studies have suggested that *c*-type cytochromes are important in electron transfer to Fe(III) in *Geobacter sulfurreducens*. However, elucidating which cytochromes are involved in Fe(III) reduction is not trivial because the genome of *G. sulfurreducens* contains genes for over 100 *c*-type cytochromes, at least 25 of which are predicted to be localized in the outer membrane where Fe(III) reduction is likely to take place. Therefore, our initial strategy has been to focus on cytochromes predicted to be localized in the outer membrane, as well as cytochromes that are specifically expressed during growth on Fe(III). Functional analysis of nearly all of the outer-membrane *c*-type cytochromes has led to the surprising result that, in many instances, the deletion of just one of the cytochrome genes severely inhibits Fe(III) reduction. This suggests that many of the multiple outer-membrane cytochromes do not serve duplicative functions, but act in concert to bring about electron transfer to Fe(III).

There are also many periplasmic cytochrome genes that are highly similar. Mutants were generated in order to evaluate their role in electron transfer to metals. It was found that PpcA, PpcB, and PpcC are required for soluble Fe(III) reduction but that mutants that could no longer produce PpcD or PpcE grew better on Fe(III) than the wild type, as did a double mutant lacking PpcB and PpcC. These results demonstrate that despite their apparent similarities in size and heme content, these periplasmic cytochromes have some different functions in electron transfer in *G. sulfurreducens* and that it is possible to make mutations that will enhance electron transfer to metals.

It has been proposed that, based on analogy to our previous findings in *Desulfovibrio vulgaris*, *c*-type cytochromes are also important electron carriers for U(VI) reduction. Analysis of over 15 *c*-type cytochrome mutants suggested that the small periplasmic *c*-type cytochromes in *G. sulfurreducens*, which are most closely related to the c_3 cytochrome responsible for U(VI) reduction in *D. vulgaris*, were not responsible for U(VI) reduction. However, knockout mutations in several outer-membrane cytochromes inhibited U(VI) reduction. These results suggest that the mechanisms for U(VI) reduction in *G. sulfurreducens* are significantly different than for *D. vulgaris* and indicate that even though U(VI) is soluble, and could potentially be reduced in the periplasm, reduction by *G. sulfurreducens* is more likely to take place primarily at the outer membrane surface.

Sequencing of the *G. sulfurreducens* genome revealed the presence of genes predicted to be involved in oxygen respiration, which was surprising because no *Geobacter* species had ever been found to grow on oxygen. However, growth conditions under which *G. sulfurreducens* can grow at oxygen concentrations that are 50% or less of atmospheric levels have now been identified. Knockout mutation studies demonstrated that growth on oxygen is dependent upon a cytochrome oxidase. The ability of *Geobacter* species to grow at low oxygen levels helps explain how they survive in aerobic subsurface environments and then rapidly respond to the development of anaerobic conditions during metals bioremediation.

Functional analysis of proteins important in central metabolism, such as a novel eukaryotic-like citrate synthase and a bifunctional succinate dehydrogenase/fumarate reductase, has also been completed. These studies are rapidly improving the understanding the physiology of *G. sulfurreducens*. This information will permit more informed decisions on strategies to optimize bioremediation and energy harvesting applications of *Geobacter* species.

19

Adapting Regulatory Strategies for Life in the Subsurface: Regulatory Systems in *Geobacter sulfurreducens*

Gemma Reguera[1]*, Cinthia Nunez[1], Richard Glaven[1], Regina O'Neil[1], Maddalena Coppi[1], Laurie DiDonato[1], Abraham Esteve-Nunez[1], Barbara Methé[2], Kelly Nevin[1], and **Derek Lovley**[1] (dlovley@microbio.umass.edu)

*Presenting author

[1]Department of Microbiology, University of Massachusetts, Amherst, MA and [2]The Institute for Genomic Research, Rockville, MD

As outlined in accompanying abstracts, *Geobacter sulfurreducens* serves as a pure culture model for the *Geobacter* species that are responsible for *in situ* uranium bioremediation in contaminated subsurface environments and that harvest electricity from waste organic matter. In order to predictively model the activity of *Geobacter*s involved in bioremediation and energy harvesting it is necessary to understand how electron transport to metals as well as central metabolism are regulated under different environmental conditions.

Of particular relevance for bioremediation and energy harvesting applications of *Geobacter* species is understanding regulation of gene expression under the sub-optimal growth conditions typically encountered in subsurface environments. For example, the genome of *G. sulfurreducens* contains a homolog of the *E. coli* stationary-phase sigma factor, RpoS, which is of interest because growth in the subsurface is likely to be analogous to the stationary phase of cultures. Survival in stationary phase, aerotolerance, growth on oxygen, and reduction of insoluble Fe(III) were diminished in an *rpoS* mutant, but there was no apparent impact on response to high temperature or alkaline pH stress, as seen in *E. coli*. In order to further elucidate the rpoS regulon, gene expression in the *rpoS* mutant and the wild type were compared with whole genome DNA microarray and proteomics approaches. These studies demonstrated that RpoS controls genes involved in Fe(III) reduction, oxygen tolerance, and oxygen respiration. This study represents the first characterization of RpoS in a member of the δ subclass of the *Proteobacteria*

and suggests that RpoS plays an important role in regulating metabolism of *Geobacter* species under the stressful conditions found in subsurface environments.

RpoS negatively regulates another sigma factor, RpoE, which modulates a distinct regulon also involved in oxygen tolerance and repair of oxidative stress damage. RpoE also was found to have an important role in controlling attachment to Fe(III) oxide and electrode surfaces, two key processes for the environmental success of *Geobacter* species. Genome-wide transcriptional profiles of *G. sulfurreducens* biofilms grown on Fe(III) oxide surfaces versus their planktonic counterparts, as well as transcriptional profiling of an *rpoE* mutant, suggested that RpoE regulates the transition from planktonic to biofilm conditions as well as maintenance of the biofilm mode of growth and electron transfer to Fe(III). These results demonstrate that RpoE and RpoS act coordinately to finely tune the adaptive responses that enable *Geobacters* to survive and outcompete many other organisms in subsurface environments.

RelA is another regulatory protein that could be important in influencing growth in the subsurface as a mutant in the putative *relA* gene in *G. sulfurreducens* grew faster than wild type under nutrient limitation. Microarray analyses of the *relA* mutant demonstrated that, as in *E. coli*, ribosomal proteins and chaperones are negatively regulated by RelA, while stress response genes are positively controlled, further suggesting that RelA may play a critical role in slow growth and stress response. In addition, RelA also appeared to positively regulate proteins required for the reduction of insoluble Fe(III) reduction, thus illustrating that in *G. sulfurreducens* RelA has unique targets that link the regulation of growth rate to metal reduction.

Analysis of the *G. sulfurreducens* genome revealed that this organism is highly attuned to its environment with 5.2% of the open reading frames in the genome dedicated to two-component proteins and 1.9% dedicated to chemotaxis. A combination of genomics and proteomics approaches identified putative histidine kinases and response regulators and results of microarray analyses of wild type and selected mutants enabled preliminary characterization and pairing of 16 two-component signal transduction proteins that previously were of unknown function and classified as "orphans". Histidine kinase knockouts in *G. sulfurreducens* over-expressing the cognate response regulators produced information on the environmental signals triggering regulatory cascades and provided further support for the role of two component systems in integrating responses to environmental stimuli with electron transfer.

Geobacter species generally live in environments high in dissolved Fe(II) and have unusually high requirements for iron due to their high cytochrome content. In *G. sulfurreducens*, concentrations of dissolved Fe(II) as high as 100 μM were found to be required for optimal growth and acetate uptake and the cellular iron content greatly exceeded that of *E. coli*, suggesting that mechanisms to regulate iron uptake and iron overload in *Geobacters* may be different than in other, previously studied organisms. A homolog of the *E. coli* Fe(II)-dependent ferric uptake regulator, Fur, was identified in the *G. sulfurreducens* genome. As in *E. coli*, expression of *fur* in *G. sulfurreducens* was repressed in the presence of Fe(II) and the phenotype of a *fur*-knockout mutant suggested that Fur has a key role in responding to changes in Fe(II) concentration in the environment. Only a small fraction of Fur-regulated genes identified by microarray analysis were preceded by a recognizable Fur box. Surprises in the genes under Fur control included proteins required for Fe(III) oxide reduction.

These studies, as well as other ongoing studies on novel regulatory strategies in *G. sulfurreducens*, suggest that models of regulation that have been developed in previ-

ously studied microorganisms can help in identifying some of the regulatory components in *G. sulfurreducens* but, in many instances, regulation patterns and mechanisms in *G. sulfurreducens* have been modified in order to adapt to life in the subsurface. Results from these regulation studies will be incorporated into the expanding *in silico* model of *G. sulfurreducens* in order to better predict the likely response of *Geobacter* species during attempts to optimize bioremediation and energy harvesting strategies.

See Also

- *In Silico* Elucidation of Transcription Regulons and Prediction of Transcription Factor Binding Sites in *Geobacter* Species Using Comparative Genomics and Microarray Clustering, Krushkal et al, on page 73.

Shewanella Federation

20

Global and Physiological Responses to Substrate Shifts in Continuous and Controlled Batch Cultures of *Shewanella oneidensis* MR-1

Jim Fredrickson (jim.fredrickson@pnl.gov), Alex Beliaev, Bill Cannon, Yuri Gorby, Mary Lipton, Peter Liu, Margie Romine, Richard Smith, and Harold Trease

Pacific Northwest National Laboratory, Richland, WA

Collaborating *Shewanella* Federation Team Leaders: Carol Giometti (Argonne NL); Eugene Kolker (BIATECH); Ken Nealson (USC); Monica Riley (MBL); Daad Saffarini (UW-M); Jim Tiedje (MSU), and Jizhong Zhou (Oak Ridge NL)

Shewanella oneidensis MR-1 is a facultative γ-Proteobacterium 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 biogeochemical cycling and in the bioremediation of contaminant metals and radionuclides. The entire genome sequence of MR-1 has been determined and high throughput methods for measuring gene expression are being developed and applied. This project is part of the *Shewanella* Federation, a multi-investigator and cross-institutional consortium formed to achieve a systems level understanding of how *S. oneidensis* MR-1 senses and responds to its environment.

Electron Acceptor Responses. To define the networks of genes responding to metal electron acceptors, mRNA expression patterns of cells reducing fumarate were compared to those reducing nitrate, thiosulfate, DMSO, TMAO, and several forms of Fe(III) and Mn(III) using whole-genome arrays of *S. oneidensis*. Analysis of variance performed on the complete dataset identified over 1600 genes displaying significant expression changes across different metal-reducing conditions. Two principal components accounted for 78% of the variability within the multiple-electron-acceptor dataset and were represented by genes displaying specific response to metals. Hierarchical clustering revealed a high degree of similarity in mRNA relative abundance levels was displayed for all the metal-reducing conditions; all clustering separately from the inorganic electron acceptors. Interestingly, no significant differences in expression profiles were observed between solid and soluble metal acceptors. Only a few genes specific for any particular metal were identified. In contrast, K-means clustering identified a group of over 150 genes displaying highly specific up-regulation under all metal-reducing conditions. Among those, we identified putative transporters, outer membrane components, as well as two electron transfer proteins (flavodoxin and a *c*-type cytochrome). Further work will be aimed at differentiating cells responses to divalent metal cations (i.e., reduction products) and the oxidized form of the metals; and functional characterization of the differentially regulated genes.

Response to O_2 Concentrations. Autoaggregation occurs in *Shewanella oneidensis* MR-1 cultures growing at high O_2 concentrations in the presence of Ca^{2+} ions. Despite the potential environmental importance of this phenomenon, little is known about the mechanisms inducing aggregate formation and subsequent impacts on cells inside the aggregates. In an effort to elucidate these mechanisms and identify processes associated with O_2-induced autoaggregation in *S. oneidensis*, a comparative analysis using DNA microarrays was performed on samples grown under different O_2 tensions in the presence and absence of Ca^{2+}. Although, when compared to O_2-limited conditions, both flocculated and unflocculated cells displayed some similarities in gene expression in response to elevated levels of O_2, including genes involved in cell envelope functions and EPS/LPS production, autoaggregation had a significant impact on gene expression in MR-1. Direct comparison of aggregated versus nonagreggated cells grown under 50% dissolved O_2 tension (DOT) revealed remarkable differences in mRNA patterns between these two states. The nonaggregated cells displayed significant increase of mRNA levels of genes involved in aerobic energy metabolism, amino acid and cofactor biosynthesis, as well as chemotaxis and motility. In contrast, genes putatively involved in anaerobic metabolism (fumarate and polysulfide reductases, and Ni/Fe hydrogenase), cell attachment (type IV pilins and curli), and transcription regulation (*rpoS*, *spo*IIAA) were upregulated under 50% DOT aggregated conditions. Notably, a gene cluster encoding outer membrane proteins and cytochromes (*mtrDEF*) also displayed up to 7-fold increase in mRNA levels in aggregated cells. Although further studies are required for resolution, we speculate that autoaggregation in *S. oneidensis* MR-1 may serve as a mechanism to facilitate reduced O_2 tensions within aggregate, leading to the expression of anaerobic genes under bulk aerobic conditions.

Carbon Metabolism. In contrast to the wide array of electron acceptors reduced by *S. oneidensis*, this organism is relatively limited in regards to utilization of multicarbon substrates for anaerobic respiration. Earlier studies indicated that MR-1 can utilize formate as a sole source of carbon and energy under anaerobic condition. A hypothetical amalgam pathway for lactate metabolism that included the elements of serine-isocitrate cycle for formate utilization was previously proposed by K. Nealson and colleagues. The availability of whole-genome sequence allowed compilation of a possible pathway for formate assimilation in *S. oneidensis.* MR-1 is predicted to possess a number of putative enzymes including pyruvate formate-lyase (PFL) that may allow for the assimilation of both exogenous and lactate-derived formate through a modification of a serine-isocitrate pathway. Three independent lines of evidence of methylotroph-like metabolism of lactate in *S. oneidensis* MR-1 are provided: lactate is stoichiometrically converted to acetate in O_2-limited or lactate-excess anaerobic chemostat cultures and no formate is present in the supernatant; the amount of PFL protein increased 2.5-fold under O_2 limited growth compared to fully aerobic growth; and relative abundance of mRNA from genes encoding key enzymes of the proposed pathway for formate assimilation including isocitrate lyase, malate synthase, and serine hydroxymethyl-transferase increased under O_2-limitation compared to aerobic growth. Moreover, biomass yield from O_2-limited or anaerobic chemostat cultures of MR-1 grown under excess lactate indicated that formate was utilized as the sole source of carbon. These results suggest the presence of an unusual mechanism of carbon metabolism of lactate in *S. oneidensis* with formate as the key element of the intermediary carbon metabolism under O_2-depleted and/or lactate excess conditions.

Protein Secretion. In many bacteria, translocation of key respiratory enzymes is mediated by the twin arginine translocation (TAT) machinery. Analysis of the N-terminal sequences of MR-1 proteins for conserved TAT leader properties revealed 30 candidates. This list includes 11 genes predicted to be responsible for utilization of

formate and H_2 as electron donors or for catalyzing the terminal step in electron transfer to nitrate, nitrite, and sulfur-containing substrates. Various uncharacterized proteases and putative redox active proteins were among the remaining 19 candidates. In order to investigate the role of secreted proteins in MR-1 respiratory metabolism, two mutants were constructed: a *tatC* gene deletion mutant and a transposon mutant of the terminal branch of type II general secretion pathway (*gspD*). The TAT mutant was unable to reduce metals with formate or H_2 as an electron donor. Although cells could grow with lactate and fumarate, there was a longer lag phase relative to the wild type. As expected, the ability to grow on nitrate and DMSO was abolished. Reduction of technetium (TcO_4^-) with lactate was severely impaired in the TAT mutant, likely due to mislocalization of one or more hydrogenases. Hydrogenase have reported to be involved in Tc reduction in other bacteria. The TAT system is a key cellular machine in MR-1 that is essential for its diverse energy metabolism.

A comprehensive proteomic approach was applied to identify candidate GSP proteins including those potentially involved in metal-reduction. Steady state O_2-limited wild type and *gspD* mutant cells of MR-1 were sampled from bioreactors for mass spectral proteomics and metal reduction activities. Proteome analysis of whole cells, cell fractions, membrane vesicles, and extracellular proteins revealed the mislocalization of several hypothetical proteins in the mutant compared to the wild type, as well as OmpW. OmpW and one hypothetical exhibited increased expression in cells incubated with various metals. The localization of MtrA, MtrB, and MtrC, proteins previously implicated in metal reduction, were unaffected according to proteome and western blot analyses. These results demonstrate that while the GSP is necessary for efficient metal reduction in these cells, several key electron transfer proteins essential for Fe(III) and Mn(IV) reduction were not mislocalized. By applying a combination of controlled cultivation integrated with proteome measurements genes that are candidate secreted proteins, including those involved in metal transformation, can be identified.

21

Integrated Analysis of Gene Functions 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], Timothy Palzkill[2], James M. Tiedje[3], Kenneth H. Nealson[4], Alex S. Beliaev[5], Ting Li[1], Xiufeng Wan[1], Steven Brown[1], Dawn Stanek[1], Weimin Gao[1], Feng Luo[1], Jianxin Zhong[1], Liyou Wu[1], Barua Soumitra[1], Crystal B. McAlvin[1], David Yang[1], Robert Hettich[1], Nathan VerBerkmoes[1], Yuri Gorby[5], Richard Smith[5], Mary Lipton[5], and James Cole[3]

[1]Oak Ridge National Laboratory, Oak Ridge, TN; [2]Baylor College of Medicine, Houston, TX; [3]Michigan State University, East Lansing, MI; [4]University of Southern California, Los Angeles, CA; and [5]Pacific Northwest National Laboratory, Richland, WA

Collaborating *Shewanella* Federation Team Leaders: Jim Fredrickson, Pacific Northwest National Laboratory, Richland, WA; Carol Giometti, Argonne National Laboratory, Argonne, IL; Eugene Kolker, BIATECH, Bothell, WA; and Monica Riley, Marine Biological Laboratory, Woods Hole, MA

Shewanella oneidensis MR-1, a facultatively anaerobic g-proteobacterium, possesses remarkably diverse respiratory capacities. In addition to utilizing oxygen as a terminal electron acceptor during aerobic respiration, *S. oneidensis* can anaerobically respire various organic and inorganic substrates, including fumarate, nitrate, nitrite, thiosulfate, elemental sulfur, trimethylamine *N*-oxide (TMAO), dimethyl sulfoxide (DMSO), Fe(III), Mn(III) and (IV), Cr(VI), and U(VI). However, the molecular mechanisms underlying the anaerobic respiratory versatility of MR-1, however, remain poorly understood. In this project, we have integrated genomic, proteomic and computational technologies to study energy metabolism of this bacterium from a systems-level perspective.

Molecular Responses to Anaerobic Growth with Different Electron Acceptors. To define the repertoire of MR-1 genes responding to different terminal electron acceptors, transcriptome profiles were examined in cells 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 grown anaerobically for 3.5 hours in the presence of different electron acceptors and compared to RNA extracted from cells grown under fumarate-reducing conditions (the reference condition). More than 1600 genes display significant expression changes across different metal-reducing conditions. Real-time PCR analysis for some selected genes showed that microarray-based quantitation is highly accurate. Hierarchical cluster analysis indicated that genes showing differential expression under metal-reducing conditions generally clustered together, whereas genes showing differences in mRNA abundance levels under non-metal respiratory conditions clustered together. Interestingly, no significant differences in expression profiles were observed between solid and soluble metal acceptors. Only a few genes specific for any particular metal were identified. In contrast, a group of over 150 genes displaying highly specific up-regulation under all metal-reducing conditions were identified, including, putative transporters, outer membrane components, as well as two electron transfer proteins (flavodoxin and a *c*-type cytochrome). In addition, a number of genes, of which 35-55% encoded hypothetical proteins, were uniquely induced or repressed in response to a single

electron acceptor. This work has yielded numerous candidates for targeted mutagenesis and represents an important step towards the goal of characterizing the anaerobic respiratory system of *S. oneidensis* MR-1 on a genomic scale.

Phage Display. Along with mass spectrometry, two-hybrid system and protein arrays, phage display is another powerful technique for studying protein-ligand interactions. The first key step of mapping protein interactions was to clone all protein-coding ORFs to allow exogeneous expression of its protein for functional analysis. We have spent tremendous efforts in cloning genes into an universal vector. Progress as of the close of 2003 is that 1,691 genes were cloned while no clones were obtained for 174 genes. Additionally, a random phage display library utilizing "shotgun" cloning of sheared *S. oneidensis* genomic DNA has been constructed.

Genetic Mutagenesis. One of the most powerful ways to define the function of a gene is to turn the gene off or change the expression by replacing the normal gene with a mutated counterpart. We have successfully modified and utilized vector systems for homologous recombination in *S. oneidensis* MR-1. Currently, our laboratory is interested in understanding transcriptional gene regulation in *S. oneidensis*. We are targeting approximately 220 annotated transcription factors (TFs) for knockout mutatgenesis. Analysis of the *S. oneidensis* genome sequence suggests that insertional mutagenesis is appropriate for only 78 of these TFs as they are transcribed in their own operon. We have also systematically knocked out some of the genes involved in metal reduction as revealed by microarray analysis discussed above.

Numerous other MR-1 genes have been successfully inactivated using a PCR-based, in-frame deletion mutagenesis strategy. Our current collection of deletion mutants includes those strains with mutations in *etrA*, *arcA*, *fur*, *crp*, *fur/etrA*, *etrA/crp*, *rpoH* (sigma-32), *ompR*, *envZ*, *oxyR*, *cya*1-3 (adenylate cyclases), and many others. Microarray-based gene expression profiling has been used to analyze a number of these mutant strains. For example, we have employed whole-genome DNA microarrays, large-scale proteomic analysis using liquid chromatography-mass spectrometry (LC-MS), and computational motif discovery tools to define the *S. oneidensis* Fur regulon. Using this integrated approach, we identified 9 probable operons (containing 24 genes) and 15 individual ORFs of either unknown function (SO0447-48-49, 0798-97, 0799, 1188-89-90, 2039, 3025, 3027, 3406-07-08, 3062, 3344, 4700, 4740) or annotated as encoding transport and binding proteins (*ftn*, *bfd*, SO1111-12, 1482, 1580, *feoAB*, *alcA*-3031-32, 3669-68-67, *tonB1-exbB1-exbD1*, *viuA*, *irgA*, 4743) that are predicted to be direct targets of Fur-mediated repression based on their up-regulated expression profiles in a *fur* deletion mutant and the presence of potential Fur-binding sites in their upstream regulatory regions. This study suggests, for the first time, a possible role of 4 operons and 8 ORFs of unknown function in iron metabolism.

Chemostat Growth Studies with MR-1 Mutant Strains. Using the growth facility at PNNL, *Shewanella oneidensis etrA* and *arcA* deletion strains and the parental strain were each grown in chemostats in continuous culture for 410 hours. The growth conditions were altered from an aerobic steady state, to a microoxic steady state and to an anaerobic steady state to examine the contribution of each regulator in *S. oneidensis*. Samples were collected at each steady state for organic acid, proteome, cytochrome and transcriptome analyses. Samples were also harvested at 0, 5, 10, 20, 30, 40, 50, 60, 90, 120, and 150 minutes after transition from aerobic to microoxic steady states for mRNA and protein analysis.

Elucidation of the Functions of a Conserved Hypothetical Protein. Whole-genome sequence analyses of a variety of microorganisms indicated that 30-60% of the identified genes encode functionally unknown proteins. Defining the functions of hypothetical proteins is a great challenge. Integrated approaches for systematic study of their functions are needed. As a first attempt, an in-frame deletion mutant was generated for the conserved hypothetical protein of 592 amino acids, SO1377. Physiological analysis showed that this mutant was very sensitive to hydrogen peroxide, showed slow growth rate under aerobic condition but not anaerobic conditions, and had higher spontaneous mutation rates. Microarray analysis revealed that numerous genes are affected by this mutation. Computational analyses of secondary and tertiary structure also revealed that the protein could have potential functions in formation of protein complexes at the inner bacterial cell membrane, ATP/GTP binding, nucleotide binding, protein transport and molecular chaperone. Overall, our results suggested this gene could be involved in iron homeostasis and oxidative damage protection in *S. oneidensis* MR-1.

Molecular Basis of Stress Responses. Other work related to Shewanella focuses on the elucidation of the molecular basis of bacterial adaptive responses to various environmental stresses, namely, heat stress, cold stress, high salt, low/high pH, oxidative stress, and metal toxicity. These studies employ primarily global gene expression profiling using cDNA/oligonucleotide microarrays and targeted gene mutagenesis. The initial manuscripts for two of these studies (heat shock and salt stress) have already been written and have been submitted for publication or are close to being submitted. In the study on oxidative stress, the effect of H_2O_2-induced oxidative stress on the gene expression profiles of *S. oneidensis* wild type and mutant strains was investigated. Microarray analysis of the wild type cells indicated significant changes in the expression levels of numerous genes that are known or have not been previously described to be involved in the oxidative stress responses of other bacterial species. Among these are the alkyl hydroperoxide reductase (*Ahp*) gene, the catalase (*Kat*) gene, the stress response DNA-binding protein (*dps*) gene, and the genes involved in the TonB transport systems. In addition, a LysR family transcriptional regulator showed immediate yet transient upregulation in response to H_2O_2 treatment, suggesting the hypothesis that it regulates H_2O_2 stress responses in *Shewanella oneidensis.* Sequence comparison and computational modeling predicted the gene to be the potential analog of the *E. coli* OxyR gene. Yet phenotype characterization of the deletion mutant of the gene revealed interesting responses toward various oxidative stimuli. Global expression profiling of the mutant indicated that the LysR regulator indeed controlled some of the genes that had been reported to belong to the OxyR regulon in *E. coli*, but it also regulated many uncharacterized genes.

In another study, we examined the response of *S. oneidensis* to high levels of heavy metals to better understand the repertoire of genes and regulatory mechanisms enabling heavy metal resistance. MR-1 was able to grow in LB medium with strontium (Sr^{2+}) concentrations as high as 180 mM, but showed substantial growth inhibition at levels above 180 mM. *S. oneidensis* resistance to 180 mM Sr was examined using DNA microarrays. Transcriptome profiles were generated from mid-exponential phase bacteria grown in the presence of Sr^{2+} and compared to profiles from MR-1 cultured to the same growth phase in the absence of strontium. The stress response of *S. oneidensis* to a shock addition of 180 mM Sr was also examined after 5, 30, 60 and 90 minutes using microarrays. Siderophore biosynthesis and iron uptake genes were highly induced (up to 622 fold) and a siderophore biosynthetic mutant was more sensitive to strontium, suggesting that siderophore production plays an integral role in the ability of *S. oneidensis* to mediate strontium resistance.

Network Modelling. Understanding the regulatory interactions between thousands of genes in a given organism from massive time-course microarray data is one of the most challenging tasks in the field of microbial functional genomics. Currently, the inference of such genetic interaction networks is harmed by the dimensionality problem because the number of genes in a genome far exceeds the number of measured time points due to high cost of measurements. It is essential to develop powerful computational tools to extract as much biological information as possible from ambiguous expression data containing noise. Different from existing methods, we are developing a computational method based on random matrix theory. We are using the matrix of pair-wise correlation to identify connections between genes. In contrast to other network identification methods, the threshold for defining network links is determined automatically and self-consistently based on the data structure itself. We have applied this method to identify regulatory networks in yeast based on the massive available microarray data. The identified gene interactions were very consistent with our knowledge, suggesting that this method is very useful for network identification. We are now further testing this method based on microarray data from Shewanella, Deinococcus, yeast, worm, fly and human.

22

Profiling *Shewanella oneidensis* Strain MR-1: Converting Hypothetical Genes into Real, Functional Proteins

Eugene Kolker (ekolker@biatech.org), Samuel Purvine, Alex F. Picone, Natali Kolker, and Tim Cherny

BIATECH, Bothell, WA

Collaborating ***Shewanella*** **Federation** Teams: **J. Fredrickson**, M. Romine, Y. Gorbi, A. Beliaev, B. Cannon (PNNL); **R. Smith, G. Anderson, K. Auberry, M. Lipton, D. Elias (PNNL); J. Tiedje**, X. Qiu, J. Cole (MSU); **K. Nealson**, S. Tsapin (USC); **M. Riley**, M. Serres (MBL); **C. Giometti**, G. Babnigg (ANL); **J. Zhou**, D. Thompson (ORNL).

Other **Collaborating** Teams: **E. Koonin**, M. Galperin, K. Makarova (NCBI); **C. Lawrence**, L.-A. McCue (WC); **B. Palsson**, A. Raghunathan, N. Price (UCSD); **H. Heffelfinger**, J. Timlin (SNL); **J. Yates**, W. Zhu (Scripps).

The progress in genome sequencing has led to a rapid accumulation in GenBank submissions of uncharacterized "hypothetical" proteins. These proteins, which have not been experimentally characterized and whose functions cannot be deduced from simple sequence comparisons alone, now comprise approximately one third of the public databases. That is, despite significant progress in the experimental research, this so called "70% hurdle" still holds, with every new genome bringing novel unknown proteins numbering in the hundreds or even thousands. Being very complex and fascinating in numerous aspects of its behavior and responses, *Shewanella oneidensis* strain MR-1 (SO) presents an even greater challenge, as over half of its predicted genes are considered hypothetical. If past performance in experimental characterization of new proteins from *Escherichia coli* K-12, roughly 25 per year, is of any predictive power, it will take many decades before the biological function of all these (SO) proteins is discovered.

Expression profiling of SO cells under multiple growth conditions done by the *Shewanella* Federation consortium was performed. Among the performed experiments are continuous and controlled batch cultures of SO cells under a variety of

different environmental conditions. These include electron acceptors and substrates, and limitations of thereof, such as O_2, Ca^{2+}, and Pi-limitations and UV-radiation stresses. Special emphasis was placed on robust, reproducible, and statistically validated results, rather than optimizing coverage of the expressed gene and protein contents for the above conditions. Earlier studies of SO presented a baseline of over 4,600 predicted genes with approximately 2,350 hypothetical ones.

SO gene profiling resulted in conservative estimation of over 4,000 expressed genes, including identification of over 1,900 hypothetical genes. Protein profiling experiments conservatively estimated approximately 1,550 expressed proteins with approximately 500 hypothetical ones. Using a combination of transcriptomic and proteomic approaches as well as statistical and computational methods, this analysis confidently identified over 450 hypothetical genes that were expressed in cells both as genes and proteins. In an attempt to understand the functions of these proteins, we used a variety of publicly available analysis tools. This resulted in exact or general functional assignments for over 200 hypothetical proteins. Accurate functional annotation of uncharacterized proteins calls for an integrative approach, combining expression studies with extensive computational analysis and curation, followed by the directed experimental verification.

23

Systems Biology of *Shewanella oneidensis* MR-1: Physiology and Genomics of Nitrate Reduction, the Radiation Stress Response, and Bioinformatics Applications

James M. Tiedje (tiedjej@msu.edu), James R. Cole, Claribel Cruz-Garcia, Joel A. Klappenbach, and Xiaoyun Qiu

Michigan State University, East Lansing, MI

Collaborating *Shewanella Federation* Team Leaders: Jim Fredrickson, Margie Romine, Yuri Gorby (PNNL); Eugene Kolker (BIATECH); and Jizhong Zhou (ORNL)

The Stress Response: Effects of Ultraviolet Radiation. Successful application of *Shewanella oneidensis* MR-1 in bioremediation applications may necessitate cellular tolerance to toxic levels of pollutants and damage-inducing radiation. Solar ultraviolet radiation (UVR) is perhaps the most mutagenic agent to which many organisms are exposed due to its abundance. We systematically investigated the stress response in MR-1 following exposure to UVC, UVB and UVA radiation. MR-1 showed extremely high sensitivity to both far- and near-UV with a D_{37} value (UVC) of 5.6% relative to *E. coli* K12. Photoreactivation conferred a significantly increased survival rate to MR-1 in both UVB and UVC irradiated cells: as much as 177- to 365-fold and 11- to 23-fold survival increase after UVC and UVB irradiation respectively. A significant UV mutability to rifampin resistance was detected in both UVC and UVB treated cells. Different gene expression profiles were observed after UVC, UVB and UVA treatments. More than 300 genes were up-regulated after UVA exposure whereas only about 100 genes were induced after UVC exposure. Although the SOS response occurred in all three treatments, the induction of key genes in the SOS regulon (e.g. *rec*A, *lex*A, *pol*B *etc*.) was most robust in response to UVC. Genes that are involved in protection from oxidative damage showed an increased expression level in both UVB and UVA treatments. Unexpectedly, we did not observe induction of genes encoding nucleotide excision repair (NER) compo-

nents (e.g. *uvr*A, *uvr*B and *uvr*D) in either UVB or UVC treatments. We were also unable to identify any potential SOS box upstream of *uvr*A, *uvr*B and *uvr*D. Complementation of *Pseudomonas aeruginosa* UA11079 (*uvr*A$^-$) with *uvr*A of MR-1 increased the UVC resistance of this strain more than three orders of magnitude, indicating the functionality of UvrA in repairing UVR-induced DNA damage. Using RT-PCR, we detected transcripts of *uvr*A, *uvr*B and *uvr*D from MR-1 in both UVR treated and untreated sample at equivalent levels, indicating that component genes of NER are constitutively expressed. Loss of the damage inducible NER system may contribute to the high sensitivity of this bacterium to UVR.

Aerobic and Anaerobic Nitrate Reduction. Nitrate is often found as a co-contaminant in metal and radionuclide contaminated groundwater, and understanding the response of *S. oneidensis* MR-1 to these compounds is critical to effective bioremediation applications. *S. oneidensis* MR-1 is capable of dissimilatory nitrate reduction to ammonia during both aerobic and anaerobic growth. Nitrate is reduced by *S. oneidensis* MR-1 in a stepwise manner from nitrate > nitrite > ammonia. Complete reduction of nitrate precedes initiation of nitrite reduction, a process controlled by thermodynamics. Genome analysis supports this physiology: *S. oneidensis* MR-1 possesses genes for a single nitrate reductase (NapA) and a single nitrite reductase (NfrA) that catalyze the reduction of nitrate to ammonia in two enzymatic steps. Expression of *napA* and *nrfA*, measured via quantitative PCR, is maximal in anaerobic batch cultures initiated with >0.5 mM nitrate. A decrease in *nap*A and *nrf*A gene expression with increasing concentrations of nitrate occurs in *E. coli*, indicating an alterative regulatory system is operating in *S. onedensis* MR-1. The expression of *nar*P/*nar*Q (NO3/NO2 sensor/response regulator) was constant in cultures fed up to 10 mM nitrate. During aerobic growth conditions with and without nitrate, *nap*A and *narP/narQ* were equivalently expressed, indicating constitutive expression of nitrate reductase activity independent of the presence oxygen or nitrate.

Nitrite accumulates during lactate-dependent nitrate reduction, and was found to inhibit growth in both aerobic and anaerobic batch culture. Decreased growth rates during chemostat culture did not alleviate nitrite toxicity due to the stepwise reduction of nitrate to ammonia. Anaerobic nitrate reduction was limited to the oxidation of lactate and pyruvate as sole carbon and energy sources - other sugars and carboxylic acids (and hydrogen) did not support growth. Growth limitation due to nitrite toxicity during aerobic batch culture was assessed using whole-genome microarrays. Significant up-regulation of genes encoding heat shock and DNA repair proteins that are associated with oxidative stress occurred in the presence of nitrite. Nitrite also resulted in the significant down-regulation of many genes involved in iron acquisition, possibly as a mechanism of reducing DNA damage induced by hydroxyl radicals generated via intracellular iron oxidation. The capacity for nitrate reduction of *S. oneidensis* MR-1 may therefore be limited by its ability to mediate oxidative damage induced by nitrite accumulation.

***MicroPlateDB* – a LIMS for Quality Control and Data Archiving Microplate Data.** In a multi-investigator research effort, such as the *Shewanella Federation*, open-access to research data and protocols is critical to genomics-level investigations involving tools such as microarrays and proteomics. We have continued development of an internet-browser accessible laboratory information management system (LIMS), the '*MicroPlateDB*', for tracking and archiving data generated during microarray construction. The LIMS is structured with the laboratory microplate as the central data type and the contents of plates are combined during virtual "reactions" as they are carried out in the. Customization of the LIMS is controlled by a set of basic data tables containing information on microplate types, contents, and

how contents of microplates are combined and stored during laboratory procedures. User-level permissions control LIMS access and allow a project manager to specify the ability to view and/or modify data on an individual basis. With a login name and password, an investigator performing microarray studies can access the LIMS to find a gene of interest in the plate used to print the array, including the concentration, size, and gel-resolved quality of the PCR-product. The user also has the ability to 'drill-down' through a set of hyperlinked microplate graphic representations to track a PCR-product from a spot on the microarray to the primers and template used to create that product. Enhancements currently under development include user-defined searching and an open-source version for public release. The LIMS was initially customized to track process information obtained during the production of a PCR-based DNA microarray for *S. oneidensis* MR-1 and has also been chosen for use in several other microarray construction projects, including the ORNL *Deinococcus radiodurans* microarray.

24

Development and Application of Optical Methods for Characterization of Protein-Protein Interactions in *Shewanella oneidensis* MR-1

Natalie R. Gassman[1*] (ngassman@chem.ucla.edu), Achillefs N. Kapanidis[1], Nam Ki Lee[1, 4], Ted A. Laurence[1,5], Xiangxu Kong[1], and **Shimon Weiss**[1,2,3]

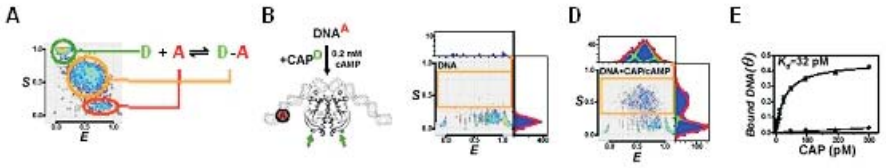
*Presenting author

[1]Dept. of Chemistry and Biochemistry, University of California, Los Angeles, CA; [2]Dept. of Physiology, David Geffen School of Medicine, University of California, Los Angeles, CA; [3]California NanoSystems Institute, University of California, Los Angeles and Santa Barbara, CA; [4]Seoul National University, Seoul, Korea; and [5]Lawrence Livermore National Laboratory, Livermore, CA

The increased availability of microbial genomes has increased the drive to elucidate the complex biological networks these microbes utilize to adapt to extreme environmental conditions. Regulation, cell adhesion, and respiration networks, which accomplish bioremediation of metals and radionuclides, are of particular interest to the Genomics:GTL program. *Shewanella oneidensis* MR-1, a dissimiliatory metal reducing bacterium with the ability to utilize a large variety of electron acceptors, is ideal for bioremediation applications. While widely studied, the mechanism by which MR-1 utilizes these electron acceptors remains unclear. Understanding these complex respiration mechanisms requires the characterization of post-translational macromolecular interactions. We are investigating these numerous protein-protein

interactions in MR-1 by developing novel optical methods for their detection and characterization.

Figure 1. Analysis of a protein-DNA interaction using ALEX-FAMS. A. E-S histogram for D-only, A-only, and D-A species with different RD-A. B. Model of CAP-DNA complex and labeling scheme. Acceptor (red) was placed on DNA, and the donor was placed on 2 possible sites on CAP (green arrows). C. E-S histogram of A-containing species, DNAA. D. E-S histogram of DNAA incubated with CAPd and 0.2 mM cAMP. E. Single molecule sorting allows calculation of biomolecular constants; ALEX-based titration of DNA with CAP in the presence (filled circles) or absence of cAMP (open circles).



We have expanded on a current single-molecule fluorescence spectroscopy (SMFS) method, single-pair Förster resonance energy transfer (spFRET), to measure stochiometry and interactions. spFRET uses a single laser to probe the transfer of excitation energy from a donor (D) fluorophore to a complementary acceptor (A) fluorophore of an interacting pair, yielding a D-A distance sensitive value *E*, which acts as a "spectroscopic ruler" for the 1-10 nm scale. While an excellent qualitative indicator of molecular interactions, the limited dynamic range of the FRET ruler precludes the measurement of interactions between large macromolecules and/or multimeric complexes. By using an alternating-laser excitation (ALEX) scheme, we have expanded the spFRET technique to report on structure, dynamics, stoichiometries, local environment and molecular interactions. This is accomplished by obtaining D-excitation and A-excitation–based observables for *ach single molecule* by rapidly alternating between D-excitation and A-excitation lasers. This scheme probes directly both FRET donors and acceptors present in a single diffusing complex and recovers distinct emission signatures for all species involved in interactions by calculating two fluorescence ratios: the FRET efficiency *E*, a distance-based ratio which reports on conformational status of the species, and a new, distance-independent stoichiometry-based ratio, *S*, which reports on the association status of the species. Two-dimensional histograms of E and *S* allow virtual sorting of single molecules by conformation and association status (Fig. 1A). ALEX is a homogeneous, "mix-and-read" assay, where interacting species are combined and optical readouts report *simultaneously* on their association status and conformational status. The potential applications of this methodology are extensive and characterization of known protein-DNA interactions, *Escherichia coli* catabolite activator protein (CAP) with DNA (Fig. 1B-D), has illustrated the method's robust nature.

The complex regulatory mechanisms governing the expression of genes involved in electron transport and energy generation in MR-1 provide a diverse array of protein-DNA and protein-protein interactions that are ideally suited for the ALEX method. One such regulatory mechanism, activated under environmental stress, is the two-component signaling cascade that initiates gene expression by the alternative sigma factor, σ^{54}. Transcriptional regulation is achieved through a cascade of protein-protein interaction that results in the interaction of a transcription regulator with the σ^{54}-RNA polymerase (RNAP) holoenzyme complex to initiate transcription. One example of this signaling cascade is the interaction of a nitrogen regulatory protein (NtrC) with σ^{54}-RNAP holoenzyme to initiate transcription of genes involved in nitrogen fixation in MR-1. Upon stimulus by environmental stress, a

sensor protein autophosphorylates resulting in the downstream phosphorylation the transcriptional regulator, NtrC. An NtrC oligomeric form then binds upstream of the promoter region and via a looped DNA intermediate catalyzes the formation of the open transcription complex. Using the ALEX methodology, we can now examine the mechanistic process of gene regulation under stress conditions from the oligomerization of the transcription regulator to the activating interaction between NtrC and the σ^{54}-RNAP holoenzyme to initiation of transcription. Progress in protein expression, site-directed mutagenesis and fluorescence labeling of MR-1 NtrC, σ^{54}-and RNAP holoenzyme will be reported.

25

Annotation of Genes and Metabolism of *Shewanella oneidensis* MR-1

Margrethe Serres and **Monica Riley** (mriley@mbl.edu)

Marine Biological Laboratory, Woods Hole, MA

Annotation

Our continuing annotation of the genes and gene products of *Shewanella oneidensis* MR-1 is taking advantage of two seldom used sources of information on protein function: (1) functions of structural domains within protein sequences, and (2) the functions of paralogous groups, groups of genes that seem to have descended from the same ancestor and tend to retain related functions.

The Structural Classification of Proteins database (SCOP) has a section describing Superfamilies of structural domains. Using a HMM method, the presence and location of structural domains has been determined for some genomes. The data for *S. oneidensis* finds structural domains in 2570 of the proteins. The data has been scrutinized in particular for all open reading frames (ORFs) having no functional assignment. 366 of unknown ORFs could be assigned some information on function from this connection.

Many bacterial genomes have genes for proteins that appear to have arisen during evolution by duplication followed by divergence. Distantly but firmly related proteins have been assembled by collecting related sequences (determined by the Darwin algorithm) into groups by transitive relationships. Such groups of sequence similar proteins can include only distantly related members in that similarity to only one protein in the group is sufficient for inclusion. No protein is a member of more than one group. Using this approach, 408 paralogous groups were identified, ranging in size from 2 to 64 members per group.

Paralogous groups give some insight to the numbers of ancestral genes required to generate contemporary bacterial gene families. Also, since paralogous group member of known function show similarity of function (sometimes closely related, sometimes more distantly), any unknown members of a paralogous group can be assigned the common denominator of function for that group.

Metabolism

To survey the metabolic capabilities or *S. oneidensis*, we recorded similarities to the protein sequences of 50 fully sequenced organisms, again using the Darwin algo-

rithm. The organisms with the largest number of "hits" were *Vibrio cholera*, *Yersinia tuberculosis*, *Escherichia coli* and *Pseudomonas aeruginosa*. Of these, biochemical information for gene products is far and away greatest for *E. coli*. Thus it was possible to assign putative enzyme function and pathway existence when homologs exist in *E. coli*, but not possible when there was similarity to a protein of unknown function in one of the other bacteria, no analogous gene in *E. coli*. Therefore broadly speaking, the metabolic capacities of *S. oneidensis* are similar to those of *E. coli*, but this is partly because more *E. coli* proteins have been characterized than in the other bacteria. Nevertheless, there were a few functions in for instance *Pseudomonas aeruginosa* such as part of the beta-ketoadipate pathway that seem to be present in *S. oneidensis* but not in *E. coli*. With further exploration of the biochemistry of organisms other than *E. coli*, more assignment of biochemical capability will be possible.

Biosynthetic capabilities for small molecule cofactors, carriers are largely intact. However in most cases where *E. coli* has two or more isozymes, *S. oneidensis* has only one. At present the broad picture for utilization of carbon sources involves very few 5 or 6 carbon sugars and sugar derivatives, rather evidence for the utilization of 3 or 2 carbon carbohydrates and organic acids. There is a defect in an essential enzyme of the glycolytic pathway. However the enzymes for utilization of some 6 carbon sugars, e.g. galactose are present and the Entner-Douderoff pathway seems to be present, completely adequate to serve 5 and 6 carbon substrates. Therefore it is not clear why many 5 and 6 carbon compounds are not utilized. This aspect bears experimental exploration.

Only some of the enzymes for reduction of organic terminal electron acceptors that are found in *E. coli* seem to be present in *S. oneidensis*. Formate metabolism is present. The well-known use of metal ions as electron acceptors could abrogate the need for using many organic acceptors. Many electron transfer intermediates are present, consistent with the unusual richness of energy transfer by this organism.

Institute for Biological Energy Alternatives

26

Estimation of the Minimal Mycoplasma Gene Set Using Global Transposon Mutagenesis and Comparative Genomics

John I. Glass[1], Nina Alperovich[1], Nacyra Assad-Garcia[1], Holly Baden-Tillson[1], Hoda Khouri[2], Matt Lewis[3], William C. Nierman[2], William C. Nelson[2], Cynthia Pfannkoch[1], Karin Remington[1], Shibu Yooseph[1], Hamilton O. Smith[1], and **J. Craig Venter**[1] (jcventer@tcag.org)

[1]Institute for Biological Energy Alternatives, Rockville, Maryland; [2]The Institute for Genomic Research, Rockville, MD; and [3]The J. Craig Venter Science Foundation Joint Technology Center, Rockville, MD

IBEA aspires to make bacteria with specific metabolic capabilities encoded by artificial genomes. To achieve this we must develop technologies and strategies for creating bacterial cells from constituent parts of either biological or synthetic origin. Determining the minimal gene set needed for a functioning bacterial genome in a defined laboratory environment is a necessary step towards our goal. For our initial rationally designed cell we plan to synthesize a genome based on a mycoplasma blueprint (mycoplasma being the common name for the class *Mollicutes*). We chose this bacterial taxon because its members already have small, near minimal genomes that encode limited metabolic capacity and complexity. We are using two mycoplasma species as platforms to develop methods for construction of a minimal cell. *Mycoplasma genitalium* is a slow-growing human urogenital pathogen that has the smallest known genome of any free-living cell at 580 kb. It has already been used to make a preliminary estimate of the minimal gene set. Global transposon mutagenesis identified 130 of the 480 *M. genitalium* protein-coding genes not essential for cell growth under laboratory conditions. That study also predicted there may be as many as 85 other *M. genitalium* genes that are similarly not essential. *Mycoplasma capricolum* subsp. *Capricolum*, an organism endemic in goats, was chosen as another platform because of its rapid growth rate and reported genetic malleability. To facilitate work with this species we sequenced and annotated its 1,010,023 bp genome. In anticipation of eventually synthesizing artificial genomes containing a minimal set of genes necessary to sustain a viable replicating bacterial cell we took two approaches to determine the composition of that gene set.

In one approach we used global transposon mutagenesis to identify non-essential genes in both of our two platform mycoplasma species. We created, isolated, and expanded clonal populations of sets of random mutants. Transposon insertion sites were determined by sequencing directly from mycoplasma genomic DNA. This effort has already expanded the previously determined list of non-essential *M. genitalium* genes, and in this study, because we isolated and propagated each mutant, we can characterize the phenotypic effects of the mutations on growth rate and colony morphology. Additionally, identification of non-essential genes in our two distantly related mycoplasma species permits a better estimate of the essential mycoplasma gene set.

In our other approach, we analyzed 11 complete and 3 partially sequenced mycoplasma genomes to define a consensus mycoplasma gene set. Previous similar

computational comparisons of genomes across diverse phyla of the eubacteria are of limited value. Because of non-orthologous gene displacement, pan-bacterial comparisons identified less than 100 genes common to all bacteria; however determination of conserved genes within the narrow mycoplasma taxon is much more instructive. The combination of comparative genomics with reports of specific enzymatic activities in different mycoplasma species enabled us to predict what elements are critical for this bacterial taxon. In addition to determining the consensus set of genes involved in different cellular functions, we identified 10 hypothetical genes conserved in almost all the genomes, and parologous gene families likely involved in antigenic variation that comprise significant fractions of each genome and presumably unnecessary for cell viability under laboratory conditions.

27

Whole Genome Assembly of Infectious ϕX174 Bacteriophage from Synthetic Oligonucleotides.

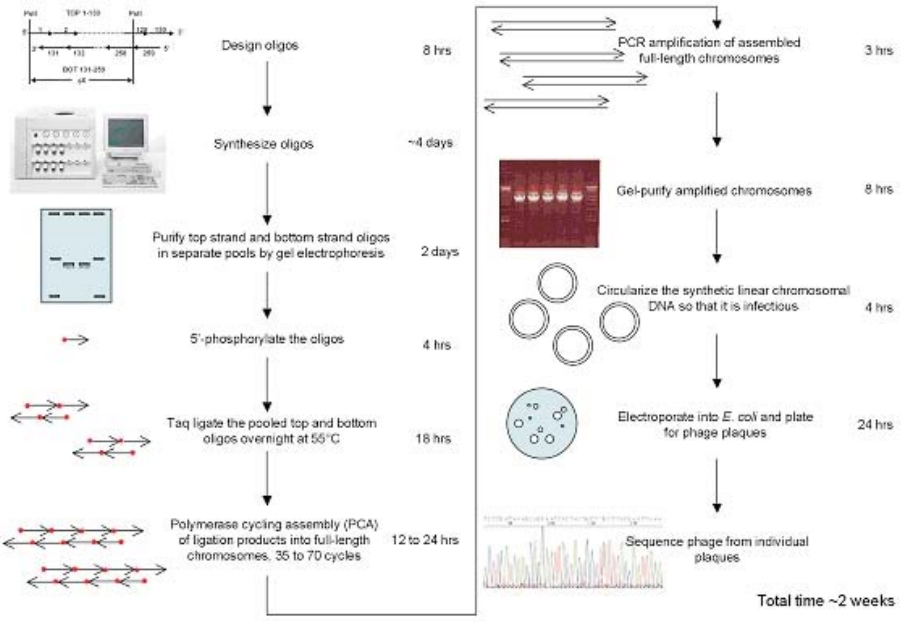
Hamilton O. Smith[1], Clyde A. Hutchison III[2], Cynthia Pfannkoch[1], and **J. Craig Venter**[1] (jcventer@tcag.org)

[1]Institute for Biological Energy Alternatives, Rockville, MD and [2]Department of Microbiology and Immunology, University of North Carolina, Chapel Hill, NC

We have improved upon the methodology and dramatically shortened the time required for accurate assembly of 5 to 6 kb segments of DNA from synthetic oligonucleotides. As a test of this methodology we have established conditions for the rapid (14 days) assembly of the complete infectious genome of bacteriophage ϕX174 (5,386 bp) from a single pool of chemically synthesized oligonucleotides. The procedure involves three key steps: 1) Gel purification of pooled oligonucleotides to reduce contamination with molecules of incorrect chain length, 2) Ligation of the oligonucleotides under stringent annealing conditions (55C) to select against annealing of molecules with incorrect sequences, and 3) Assembly of ligation products into full length genomes by polymerase cycling assembly (PCA), a non-exponential reaction in which each terminal oligonucleotide can be extended only once to produce a full-length molecule. We observed a discrete band of full-length assemblies upon gel analysis of the PCA product, without any PCR amplification. PCR amplification was then used to obtain larger amounts of pure full-length genomes for circularization and infectivity measurements. The synthetic DNA had a lower infectivity than natural DNA, indicating approximately one lethal error per 500 bp. However, fully infectious ϕX174 virions were recovered following electroporation into *E. coli*. Sequence analysis of several infectious isolates verified the accuracy of these synthetic genomes. One such isolate had exactly the intended sequence. We propose to assemble larger genomes by joining separately assembled 5 to 6 kb segments; approximately 60 such segments would be required for a minimal cellular genome. Below is a schematic diagram of the steps in the global synthesis of infectious ϕX174 bacteriophage from synthetic oligonucleotides.

The power of the above global assembly method will be fully realized when methods to remove errors from the final product are developed. Further experiments are underway to increase the efficiency of error correction.

Fig. 1. Schematic diagram of the steps in the global synthesis of infectious ϕX174 bacteriophage from synthetic oligonucleotides.



28

Development of a *Deinococcus radiodurans* Homologous Recombination System

Sanjay Vashee, **Ray-Yuan Chuang**, Christian Barnes, Hamilton O. Smith, and **J. Craig Venter** (jcventer@tcag.org)

Institute for Biological Energy Alternatives, Rockville, MD

A major goal of our Institute is to rationally design synthetic microorganisms that are capable of carrying out the required functions. One of the requirements for this effort entails the packaging of the designed pathways into a cohesive genome. Our approach to this problem is to develop an efficient in vitro homologous recombination system based upon *Deinococcus radiodurans* (Dr). This bacterium was selected because it has the remarkable ability to survive 15,000 Gy of ionizing radiation. In contrast, doses below 10 Gy are lethal to almost all other organisms. Although hundreds of double-strand breaks are created, Dr is able to accurately restore its genome without evidence of mutation within a few hours after exposure, suggesting that the bacterium has a very efficient repair mechanism. The major repair pathway is thought to be homologous recombination, mainly because Dr strains containing mutations in *recA*, the bacterial recombinase, are sensitive to ionizing radiation.

Since the mechanism of homologous recombination is not yet well understood in Dr, we have undertaken two general approaches to study this phenomenon. First, we are establishing an endogenous extract that contains homologous recombination activity. This extract can then be fractionated to isolate and purify all proteins that perform homologous recombination. We are also utilizing information from the sequenced genome. For example, homologues of *E. coli* homologous recombination proteins, such as recD and ruvA, are present in Dr. Thus, another approach is to assemble the homologous recombination activity by purifying and characterizing the analogous recombinant proteins. However, not all genes that play a major role in homologous recombination have been identified by annotation.

As a case in point, there are two candidates for the single-stranded DNA binding protein, Ssb (Dr0099 and Dr0100). To determine which of the two is the real Ssb, we first resequenced the Ssb region. We discovered two single-base deletions that when corrected give rise to a contiguous gene that contains two Ssb OB fold domains. We have purified the recombinant protein almost to homogeneity and characterized its DNA binding and strand-exchange properties. Our results suggest that despite some minor differences, the *Deinococcus* Ssb is very similar to the *E. coli* protein. In addition, using antibodies we have raised against DrSsb, we have determined that the amount of DrSsb protein, like recA, increases in the cell when exposed to a DNA damaging agent.

29

Environmental Genome Shotgun Sequencing of the Sargasso Sea

J. Craig Venter (jcventer@tcag.org), Karin Remington, Jeff Hoffman, Holly Baden-Tillson, Cynthia Pfannkoch, and Hamilton O. Smith

Institute for Biological Energy Alternatives, Rockville, MD

We have applied whole genome shotgun sequencing to pooled environmental DNA samples in this study to test whether new genomic approaches can be effectively applied to gene and species discovery and to overall environmental characterization. To help ensure a tractable pilot study, we sampled in the Sargasso Sea, a nutrient-limited, open ocean environment. Further, we concentrated on the genetic material captured on filters sized to isolate primarily microbial inhabitants of the environment, leaving detailed analysis of dissolved DNA and viral particles on one end of the size spectrum, and eukaryotic inhabitants on the other, for subsequent studies.

Surface water samples were collected from three sites off the coast of Bermuda in February 2003. Additional samples were collected from a neighboring fourth site in May 2003. Genomic DNA was extracted from filters of 0.1 to 3.0 microns, and genomic libraries with insert sizes ranging from 2-6kb were made and sequenced from both ends. The 1.66 million sequences from the February samples were pooled and assembled to provide a single master assembly for comparative purposes. An additional 325,608 reads from the May samples were also analyzed. The assembly generated 64,398 scaffolds ranging in size from 826 bp to 2.1 Mbp, containing 256 Mbp of unique sequence and spanning 400 Mbp. Evidence-based gene finding revealed 1,214,207 genes within this dataset, including 1412 distinct small subunit rRNA genes. With this set of rRNA genes, using a 97% sequence similarity cut-off to distinguish unique phylotypes, we identified 148 novel phylotypes in our

sample when compared against the RDP II database[2]. Because the copy number of rRNA genes varies greatly between taxa (more than an order of magnitude among prokaryotes), rRNA-based phylogeny studies can be misleading. Therefore, we constructed phylogenetic trees using various other represented phylogenetic markers found in our dataset. Assignment to phylogenetic groups shows a broad consensus among the different phylogenetic markers.

Just as phylogenetic classification is strengthened by a more comprehensive marker set, so too is the estimation of species richness. In this analysis, we define "genomic" species as a clustering of assemblies or unassembled reads more than 94% identical on the nucleotide level. This cut-off, adjusted for the protein-coding marker genes, is roughly comparable to the 97% cut-off traditionally used for rRNA. Thus-defined, the mean number of species at the point of deepest coverage was 451; this serves as the most conservative estimate of species richness. However, in most of the samples we observed an average maximum abundance of only 3.3%. This is a level of diversity akin to what has been observed in terrestrial samples[3].

While counts of observed species in a sample are directly obtainable, the true number of distinct species within a sample is almost certainly greater than that which can be observed by finite sequence sampling. Modeling based on assembly depth of coverage indicates that there are at least 1,800 species in the combined sample, and that a minimum of 12-fold deeper sampling would be required to obtain 95% of the unique sequence. Further, the depth of coverage modeling is consistent with as much as 80% of the assembled sequence being contributed by organisms at very low individual abundance, compatible with total diversity orders of magnitude greater than the lower bound just given. The assembly coverage data also implies that more than 100Mbp of genome (i.e., probably more than 50 species) is present at coverage high enough to permit assembly of a complete or nearly-complete genome were we to sequence to 5- to 10-fold greater sampling depth.

We demonstrate the utility of such a dataset with a study of genes relevant to photobiology within the Sargasso Sea. The recent discovery of a homolog of bacteriorhodopsin in an uncultured γ-proteobacteria from the Monterey Bay revealed the basis of a novel form of phototrophy in marine systems[4] that was observed previously by oceanographers[5,6]. Environmental culture-independent gene surveys with PCR, have since shown that proteorhodopsin is not limited to a single oceanographic location, and revealed some 67 additional closely related proteorhodopsin homologs[7]. More than 782 rhodopsin homologs were identified within our dataset, increasing the total number of identified proteorhodopsins by almost an order of magnitude. In total, we have identified 13 distinct subfamilies of rhodopsin-like genes. These include four families of proteins known from cultured organisms (halorhodpsin, bacterorhodopsin, sensory opsins, and fungal opsin), and 9 families from uncultured species of which 7 are only known from the Sargasso Sea populations.

While we are a long way from a full understanding of the biology of the organisms sampled here, even this relatively small study demonstrates areas where important insights may be gained from the comprehensive nature of this approach. Our assembly results demonstrate one can apply whole-genome assembly algorithms successfully in an environmental context, with the only real limitation being the sequencing cost.

References

1. The authors acknowledge the significant contributions of their collaborators on this project: J. Heidelberg, J.A. Eisen, D. Wu, I. Paulsen, K.E. Nelson, W. Nelson, D. E. Fouts, O. White and J. Peterson at The Institute for Genomic Research, A.L. Halpern, D. Rusch, and S.l Levy at The Center for the Advancement of Genomics, A. H. Knap, M. W. Lomas and R. Parsons at the Bermuda Biological Station for Research, Y. Rogers at the JCVSF Joint Technology Center, and K. Nealson at the University of Southern California.

2. J. R. Cole et al., *Nucleic Acids Research* **31**, 442 (Jan 1, 2003).

3. T. P. Curtis, W. T. Sloan, J. W. Scannell, *Proceedings of the National Academy of Sciences of the United States of America* **99**, 10494 (AUG 6, 2002).

4. O. Beja et al., *Science* **289**, 1902 (Sep 15, 2000).

5. Z. S. Kolber, C. L. Van Dover, R. A. Niederman, P. G. Falkowski, *Nature* **407**, 177 (Sep 14, 2000).

6. Z. S. Kolber et al., *Science* **292**, 2492 (Jun 29, 2001).

7. G. Sabehi et al., *Environ Microbiol* **5**, 842 (Oct, 2003).

Communication

30

Communicating Genomics:GTL

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

Genome Management Information System, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN

Accelerating GTL Science

For the past 15 years, the Human Genome Management Information System (HGMIS) has focused on presenting Human Genome Project (HGP) information and on imparting knowledge to a wide variety of audiences. Our goal has been to help ensure that investigators could participate in and reap the scientific bounty of this revolution, new generations of students could be trained, and the public could make informed decisions regarding complicated genetics issues. Since 2000, HGMIS has built on this experience to communicate about the DOE Office of Science's Genomes to Life program, sponsored jointly by the Office of Biological and Environmental Research (BER) and the Office of Advanced Scientific Computing Research (OASCR). To reflect our mission more accurately, HGMIS is changing its name to Genome Management Information System, and Genomes to Life recently became the Genomics: GTL systems biology program.

Genomics: GTL (GTL) is a departure into a new territory of complexity and opportunity requiring contributions of interdisciplinary teams from the life, physical, and computing sciences and necessitating an unprecedented integrative communications approach. Because each discipline has its own perspective and language, effective communication, in addition to technical achievement, is highly critical to the overall coordination and success of GTL. Part of the challenge is to help groups speak the same language, from team and research-community building and strategy development through program implementation and the reporting of results to technical and lay audiences. Our mission is to inform and foster participation by the greater scientific community and administrators, educators, students, and the general public.

Specifically, our goals center on accelerating GTL science and subsequent applications. They include the following:

- Foster information sharing, strategy development, and communication among scientists and across disciplines to accomplish synergies, innovation, and increased integration of knowledge. Emerging from this effort will be a new research community centered around the advanced concepts in GTL.
- Help reduce duplication of effort.

- Increase public awareness about the importance of understanding microbial systems and their capabilities. This information is critical not only to DOE mission needs in energy and environment but to the international community as well.

In our work with interdisciplinary teams assembled by BER and OASCR to discuss and develop scientific and programmatic strategies for accelerating the progress of GTL, we create internal documentation Web sites that organize draft texts, presentations, graphics, supplementary materials, and links. From such team activity arose a number of important documents, including more than 30 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 books.
- HGP to GTL transitional poster for the public.

We are working with DOE staff and teams of scientists to develop the next program and facilities roadmap for GTL. This roadmap, a planning and program management tool, will be reviewed by the National Academy of Sciences. GTL facilities are part of the Office of Science director's 20-year plan for frontier research facilities that will become part of the national science infrastructure.

All GTL publications are on the public Web site, which 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 15 million hits per month. Over a million text-file hits from more than 300,000 user sessions last about 13 minutes—well above the average time for Web visits. We are leveraging this Web activity to increase visibility for the GTL program.

For outreach and to increase program input and grantee base, we also identify 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, IEEE Society, National Science Teachers Association, National Association for Biology Teachers, and Biotechnology Industry Organization (BIO), as well as at carbon sequestration meetings and the G8 energy ministers' conference. We also organize "Meet the Funders" and special GTL presentations at national and international meetings of BIO and ASM. We mail some 1600 packages of educational material each month to requestors, and we furnish handouts in bulk to meeting organizers who are hosting genomics educational events.

In the past year, we participated in the closing of the HGP and the accompanying exhibition in Congress. We continue to create and update handouts, including a new primer that explores the impact of genomics on science and society, as well as flyers on careers in genetics and on relevant issues of concern to minority communities. We supply educational materials in print and on the Web site about ethical, legal,

and social issues (called ELSI) surrounding the increased availability of genetic information.

We helped draft the ocean ecogenomics sensing concept being developed by NOPP, a confederation of 15 federal agencies (including DOE) that seeks to provide leadership and coordination of national oceanographic research and education programs. Ecogenomics is a new field that increasingly will be empowered by the results of GTL and other programs.

In anticipation of communications needs and new avenues to more comprehensively represent GTL science to multidisciplinary audiences, 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 and Office of Advanced Scientific Computing 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

31

Global Organization of Metabolic Fluxes in the Bacterium, *Escherichia coli*

E. Almaas[1], B. Kovács[1,2], T. Vicsek[2], Z. N. Oltvai[3] and **A.-L. Barabási**[1] (alb@nd.edu)

[1]Department of Physics, University of Notre Dame, Notre Dame, IN; [2]Biological Physics Department and Research Group of HAS, Eötvös University, Budapest, Hungary; and [3]Department of Pathology, Northwestern University, Chicago, IL

Cellular metabolism, the integrated interconversion of thousands of metabolic substrates through enzyme-catalyzed biochemical reactions, is the most investigated complex intercellular web of molecular interactions. While the topological organization of individual reactions into metabolic networks is increasingly well understood, the principles governing their global functional utilization under different growth conditions pose many open questions. We have implemented a flux balance analysis (FBA) of the *E. coli* MG1655 metabolism, finding that the network utilization is highly uneven: while most metabolic reactions have small fluxes, the metabolism's activity is dominated by several reactions with very high fluxes[1]. *E. coli* responds to changes in growth conditions by reorganizing the rates of selected fluxes predominantly within this high flux backbone. The identified behavior likely represents a universal feature of metabolic activity in all cells, with potential implications to metabolic engineering.

To identify the interplay between the underlying topology[2,3] of the *E. coli* K12 MG1655 metabolic network and its functional organization, we focused on the global features of potentially achievable flux states in this model organism with a fully sequenced and annotated genome. In accordance with FBA[4-7], we first identified the solution space (i.e., all possible flux states under a given condition) using constraints imposed by the conservation of mass and the stoichiometry of the reaction system for the reconstructed *E. coli* metabolic network. Assuming that cellular metabolism to be in a steady state and optimized for the maximal growth rate, FBA allows us to calculate the flux for each reaction using linear optimization, providing a measure of each reaction's relative activity. A striking feature of the obtained flux distribution[1] is its overall inhomogeneity: reactions with fluxes spanning several orders of magnitude coexist under the same conditions. To characterize the coexistence of such widely different flux values, we plot the flux distribution for active (non-zero flux) reactions of *E. coli* grown in a glutamate- or succinate-rich substrate. The distribution is best fitted with a power law with a small flux constant, indicating that the probability that a reaction has flux v follows $P(v) \sim (v + v_0)^{-\alpha}$, where the constant is $v_0 = 0.0003$ and the flux exponent has the value $\alpha = 1.5$. The observed power-law is consistent with published experimental data as well[1,8].

We further examined whether these observed flux distributions are independent of the exocellular conditions by mimicking the influence of various growth conditions by randomly choosing 10%, 50% or 80% of the 96 potential substrates that *E. coli*

can consume in this *in silico* model. Optimizing the growth rate, we find that the power law distribution of metabolic fluxes is in fact independent of the external conditions. Moreover, the implementation of a "hit-and-run" method, which samples the solution space in 50,000 non-optimal states, confirms that the power law flux distribution also is independent of the assumption of optimality[1].

The observation and theoretical prediction of a power-law load distribution in simple models, as well as the presence of a power law in both the optimal and non-optimal flux states, suggests that the metabolic flux organization is a direct consequence of the network's scale-free topology. As all organisms examined to date are characterized by a scale-free metabolic network topology, the observed scaling in the flux distribution is likely not limited to *E. coli*, but characterizes all organisms from eukaryotes to archaea. As FBA is available for an increasing number of prokaryotic- and eukaryotic organisms, this prediction could be verified both experimentally and theoretically in the near future. Hence, the observed uneven local and global flux distribution appears to be rooted in the subtle, yet generic, interplay of the network's directed topology and flux balance, channeling the numerous small fluxes into high flux pathways. The dependence of the scaling exponents characterizing the flux distributions on the nature of the optimization process, as well as the experimentally observed exponent, may serve as a benchmark for future structural and evolutionary models aiming to explain the origin, the organization and the modular structure of cellular metabolism.

References

1. E. Almaas, B. Kovács, T. Vicsek, Z. N. Oltvai and A.-L. Barabási, *Nature*, in press.
2. H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai and A.-L. Barabási, *Nature* **407**, 651-4 (2000).
3. E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai and A.-L. Barabási, *Science* **297**, 1551-5 (2002).
4. J.S. Edwards and B.O. Palsson, *Proc Natl Acad Sci U S A* **97**, 5528-33 (2000).
5. J.S. Edwards, R.U. Ibarra, and B.O. Palsson, *Nat Biotechnol* **19**, 125-30 (2001).
6. R.U. Ibarra, J.S. Edwards and B.O. Palsson, Nature 420, 186-9 (2002).
7. D. Segre, D. Vitkup and G.M. Church, *Proc Natl Acad Sci U S A* **99**, 15112-7 (2002).
8. M. Emmerling et al., *J Bacteriol* **184**, 152-64 (2002).

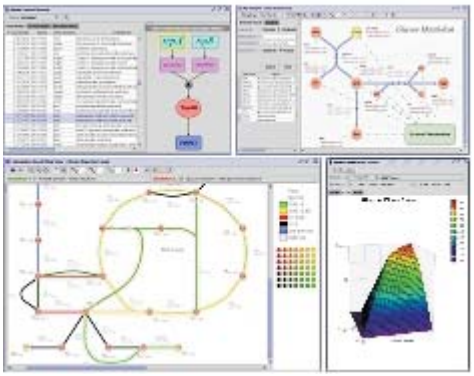
32

SimPheny™: Establishing a Computational Infrastructure for Systems Biology

Christophe H. Schilling (cschilling@genomatica.com), Sean Kane, Martin Roth, Jin Ruan, Kurt Stadsklev, Rajendra Thakar, Evelyn Travnik, and Sharon Wiback

Genomatica, Inc., San Diego, CA

The Genomics:GTL (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 constraint-based modeling approach.



We are currently utilizing this platform for a number of DOE-related projects that are discussed in accompanying posters and abstracts including:

1. Analysis and Design of Genome-scale Metabolic Networks (co P.I. Christophe Schilling, Costas Maranas, Penn State University)

2. *In Silico* Modeling to Improve Uranium Bioremediation and Energy Harvesting by *Geobacter* species (P.I. Derek Lovley, University of Massachusets, Amherst)

3. Development and Application of a Genome-scale Metabolic Model for *Pseudomonas fluorescens* (P.I. Sung Park, Genomatica, Inc.)

Recently we have launched another SBIR research program to address the problem of how to effectively deploy and deliver a system such as SimPheny to the academic research community to effectively promote collaborative research around systems biology. Ultimately we seek to establish an academic/institutional access program for the distribution, support, and training of our systems biology software platform to enable a broader usage of model-driven research for enhanced biological discovery. To accomplish this we are systematically addressing key issues related to deployment

strategies, collaborative requirements, experimental data integration needs, as well as modeling and simulation requirements. This research will be accomplished by working with a number of existing collaborators and groups involved with the DOE Genomics:GTL program that represent different types of user groups. Collectively, success with this program will facilitate the research activities of laboratories involved in various microbial genome programs and provide a much-needed solution to their data integration needs through the introduction of model centric databases. The results of these research activities will also provide valuable information on the collaborative needs and system requirements for the development of complementary software platforms that may be under parallel development by other groups. Perhaps most importantly, success with this program will further one of the core aims of the Genomics:GTL program, namely the development and distribution of a computational infrastructure for systems biology research. Establishing such an institutional/academic technology access program will also enable Genomatica to distribute and license non-energy related microbial models to the general scientific community for applications related to both medical and industrial biotechnology.

33

Analysis and Design of Genome-Scale Metabolic Networks

Costas D. Maranas[1] (costas@psu.edu), Anthony P. Burgard[1], Evgeni V. Nikolaev[1], Priti Pharkya[1], and Christophe H. Schilling[2]

[1]Department of Chemical Engineering, Pennsylvania State University, University Park, PA and [2]Genomatica, Inc., San Diego, CA

An overarching attribute of metabolic networks is their inherent robustness and ability to cope with ever changing environmental conditions. Despite this flexibility, network stoichiometry and connectivity do establish limits/barriers to the coordination and accessibility of reactions. The recent abundance of complete genome sequences has enabled the generation of genome-scale metabolic reconstructions for various microorganisms(1,2). Here we introduce the Flux Coupling Finder (FCF) framework for elucidating the topological and flux connectivity features of genome-scale metabolic networks(3). The framework is demonstrated on genome-scale metabolic reconstructions of Helicobacter pylori, Escherichia coli, and Saccharomyces cerevisiae(4-6). The analysis allows one to determine if any two metabolic fluxes, v_1 and v_2, are (i) directionally coupled, if a non-zero flux for v_1 implies a non-zero flux for v_2 but not necessarily the reverse; (ii) partially coupled, if a non-zero flux for v_1 implies a non-zero, though variable, flux for v_2 and vice-versa; or (iii) fully coupled, if a non-zero flux for v_1 implies not only a non-zero but also a fixed flux for v_2 and vice-versa. Flux coupling analysis also enables the global identification of blocked reactions, which are all reactions incapable of carrying flux under a certain condition, equivalent knockouts, defined as the set of all possible reactions whose deletion forces the flux through a particular reaction to zero, and sets of affected reactions denoting all reactions whose fluxes are forced to zero if a particular reaction is deleted. The FCF approach thus provides a novel and versatile tool for aiding metabolic reconstructions and guiding genetic manipulations.

The advent of genome-scale metabolic models has also laid the foundation for the development of computational procedures for suggesting genetic manipulations that lead to overproduction. Here the computational OptKnock framework is introduced for suggesting gene deletions strategies leading to the overproduction of

chemicals or biochemicals in *E. coli*(7,8). This is accomplished by ensuring that a drain towards growth resources (i.e., carbon, redox potential, and energy) must be accompanied, due to stoichiometry, by the production of a desired product. Computational results for gene deletions for succinate, lactate, and 1,3-propanediol (PDO) production are in good agreement with mutant strains published in the literature. While some of the suggested deletion strategies are straightforward and involve eliminating competing reaction pathways, many others suggest complex and non-intuitive mechanisms of compensating for the removed functionalities. The OptKnock procedure, by coupling biomass formation with chemical production, hints at a growth selection/adaptation system for indirectly evolving overproducing mutants.

References

1. J. L. Reed and B. O. Palsson (2003). "Thirteen years of building constraint-based in silico models of Escherichia coli." *J Bacteriol* **185**(9): 2692-9.

2. M. W. Covert, C. H. Schilling, I. Famili, J. S. Edwards, I. I. Goryanin, E. Selkov and B. O. Palsson (2001). "Metabolic modeling of microbial strains in silico." *Trends Biochem Sci* **26**: 179-186.

3. A. P. Burgard, E. V. Nikolaev, C. H. Schilling and C. D. Maranas (2004). "Flux coupling analysis of genome-scale metabolic network reconstructions." *Genome Res*, in press.

4. C. H. Schilling, M. W. Covert, I. Famili, G. M. Church, J. S. Edwards and B. O. Palsson (2002). "Genome-scale metabolic model of Helicobacter pylori 26695." *J Bacteriol* **184**(16): 4582-93.

5. J. S. Edwards and B. O. Palsson (2000). "The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities." *Proc Natl Acad Sci U S A* **97**(10): 5528-33.

6. J. Forster, I. Famili, P. Fu, B. O. Palsson and J. Nielsen (2003). "Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network." *Genome Res* **13**(2): 244-53.

7. P. Pharkya, A. P. Burgard and C. D. Maranas (2003). "Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock." *Biotechnol Bioeng* **84**: 887-899.

8. A. P. Burgard, P. Pharkya and C. D. Maranas (2003). "Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization." *Biotechnol Bioeng* **84**(6): 647-57.

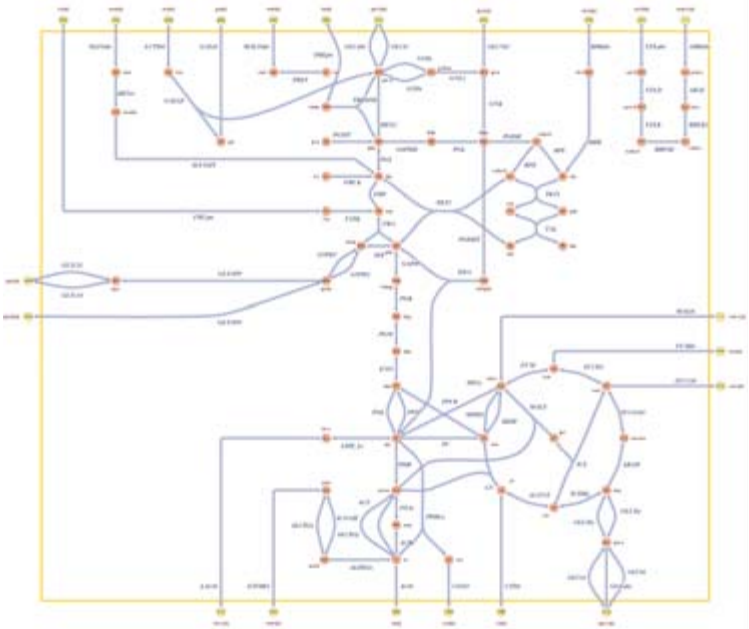
34

Development and Industrial Bioprocessing Application of a Genome-Scale Metabolic Model for *Pseudomonas fluorescens*

Sung M. Park[1] (spark@genomatica.com), Christophe H. Schilling[1], Tom Ramseier[2], and Charles Squires[2]

[1]Genomatica, Inc., San Diego, CA and [2]Dow Chemical Company

Innovative approaches are needed to utilize the information generated from genome research in an integrated fashion to analyze, interpret, and predict the function of biological systems and assist the advancement of biotechnology on the whole. This work addresses these needs with novel engineering approaches for studying the systemic capabilities of metabolism in completely sequenced bacterial genomes. The overall goal of this entire SBIR research program is to demonstrate the utility of constraints-based modeling to drive metabolic engineering and the design of bioprocesses utilizing *Pseudomonads*. There are two main commercial applications for altering the metabolism of these organisms, which include their use as a catalyst for the fermentative production of various biologics (e.g. industrial enzymes, and chemicals) as well as their use in bioremediation treatment strategies. In collaboration with the Dow Chemical Company, we have been addressing the commercial needs to fully implement the model for metabolic engineering objectives. The plan represents an integrated effort including computational and experimental components along with the necessary software development required to support these efforts.



This poster focuses exclusively on the development and implementation of a genome-scale metabolic model of *Pseudomonas fluorescens* that we have accomplished through our SBIR Phase I effort. This model is now the subject of further enhancement and utilization in a Phase II program currently underway. A comprehensive *in*

silico metabolic model of *P. fluorescens* will be shown within SimPheny. Metabolic model reconstruction of *P. fluorescens* was primarily based on the genome sequence with additional information obtained from the literature. The model includes all of the major metabolic pathways in this organism and contains 928 balanced chemical reactions accounting for 1244 genes (~20% of the total genes in *P. fluorescens*). Simulations with the reconstructed model show a range of metabolites that can be taken up and be degraded.

Modeling and simulation strategies for systems biology can now be used to guide experimental design, facilitate biological discovery, and produce the next generation of enhancements to metabolism-dependent bioprocesses. This model of *P. fluorescens* provides another platform to demonstrate power of genome-enabled science and the potential for using modeling technology to drive biological research.

35

Parallel Scaling in Amber Molecular Dynamics Simulations

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

Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA

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 decomposition" 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.

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. 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 exploring several methods of reducing the communications costs of PME in highly parallel systems.

Current Code Status

Two separate implementations of spatial decomposition are in the final testing stage, and will be released in March, 2004, as a part of Amber 8. The first, called *psander*, builds upon "classic" Amber code, and promises to minimize communication times, particularly on systems such as clusters of relatively low-end machines. The second, called *pmemd*, is in some ways a more ambitious effort: it involves an extensive re-write of major portions of the code in a controlled F90 environment, carefully moving subsets of features in as they can be validated. *Pmemd* is best suited for large numbers of processors that have good communications; for example, it scales well to 128 or 256 processors on systems such as IBM SPx architectures or the Lemiuex supercomputer at the Pittsburgh Supercomputer Center. Timings, capabilities, and prospects for future development will be presented in our poster.

References

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).

36

Bioinformatics Methods for Tandem Mass Spectrometry

Andrey Gorin[1] (agor@ornl.gov), Tema Fridman[1], Robert M. Day[1], Jane Razumovskaya[2], Andrei Borziak[1], and Edward Uberbacher[2]

Computational Biology Institute, [1]Computer Science and Mathematics Division, [2]Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN

The importance of computational tools and algorithms for mass spectrometry (MS) is hard to overstate. Efficient and reliable software for database search identification of proteins is the foundation of today high-throughput protein identification based on mass spectrometry data. At the same time unrelenting pace of scientific research creates a constant demand for a higher precision, efficiency and novel capabilities of the bioinformatics and computational algorithms used to analyze MS data.

We are developing a set of novel computational algorithms for reliable and comprehensive protein identification through detailed analysis of the tandem MS (MS/MS) data. The principal idea could be described as "micro analysis" of the spectra: analysis of the patterns typical for individual peak categories and relationships between individual peaks. Our approach is to design a probabilistic classification algorithm, aimed to establish identities of individual peaks in terms of belonging to the specific peak categories. A large set of positively identified peptide spectra have been used to determine "neighborhood" patterns for b- and y-ions as well as conditional probabilities of other important observable attributes for peak categories of interest. The established patterns have been applied to determine peak identities in other tandem MS spectra. The identification is done in a probabilistic manner, so the results have the form of probabilistic statements (e.g., "peak number 123 is a b-ion with a 0.8 probability"). The robustness of the method (named Probability Profile Method (PPM)) was investigated on a large set (>5000) of positively verified peptide spectra. Preliminary results indicate that a large majority of the useful peaks in MS/MS spectra could be identified with a surprising level of confidence, providing founda-

tion for a range of new algorithmic capabilities. An incomplete of the possible directions includes: (1) spectra can be edited sorting out desirable peak categories; (2) overall characteristics of MS/MS spectra, such as parent ion charge or total number of the present useful b- and y-ions, can be very rapidly estimated with a high precision; (3) labeled peaks of the same category, e.g. b-ion peaks, can be connected into *de novo* peptide tags providing a way for protein identification without strong reliance on the sequence database.

Two specific applications of the PPM algorithm will be discussed in details.

First, we report a novel tool for differentiation of parent ion charge states. For each spectrum we predict number of the fragments ions with charges 1 and 2. Spectra of the parent charge 3 have those fragments roughly equally 1:1, as one would intuitively expect with a splitting of +3 charge. At the same time the 2++ parent ion has 7-fold more single charged fragments compare to double charged. We demonstrate that the total number for each type is very accurately computed without any prior assumptions about what parent charge state is. As a result the PPM-based tool is fast and has 99% accuracy while being applicable to a wide range of peptide spectra. Importantly the parent charge differentiation capability not only to 2-times acceleration of the identification process, but also may eliminate some hard-to-catch misidentification originating from the wrong parent mass estimate.

Second, we demonstrate a PPM-based approach to construction of de novo peptide tags. Efficient separation of "noble" b- and y-ions dramatically simplifies algorithmic challenges, as we can easily generate and score all connectable paths for de novo tags, without "prefixing" or elaborated optimization techniques. Our method is capable of finding peptide tags 3 to 10 amino acid long for ~80% of MS/MS spectra from our testing set. When only a single top scoring tag was considered for the answer, more than a half of the constructed tags were correct ones. While additional tests and development are needed before de novo sequencing could be declared a solved problem, the approach holds a strong promise to substantially improve performance of several bioinformatics tools depending on de novo methodology for MS/MS data analysis.

This work was funded in part by the US Department of Energy's Genomics:GTL program (www.doegenomestolife.org) under two projects, "Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling" (www.genomes-to-life.org) and "Center for Molecular and Cellular Systems" (www.ornl.gov/GenomestoLife).

37

The Use of Microarray Technology and Data Mining Techniques to Predict Gene Regulation and Function in *Geobacter sulfurreducens*

Barbara Methé[1]* (bmethe@tigr.org), Kelly Nevin[2], Jennifer Webster[1], and **Derek Lovley**[2]

*Presenting author

[1]The Institute for Genomic Research, Rockville, MD and [2]University of Massachusetts, Amherst, MA

Geobacter sulfurreducens is a member of a family of prokaryotes which possess the ability to oxidize organic compounds to carbon dioxide with Fe(III) or other metals serving as the electron acceptor. As such they play critical roles in the global cycles of these metals and carbon. Additional interest in *Geobacter* spp. stems from their potential as agents of bioremediation via an ability to precipitate soluble metals including uranium and a capacity to create electricity that can be captured via energy-harvesting electrodes. Completion of the entire *G. sulfurreducens* genome sequence has provided the critical foundation for the creation of a whole genome microarray to examine global gene expression patterns.

Two categories of experiments for querying whole genome PCR-based arrays are currently being pursued: 1) test wild type *G. sulfurreducens* gene expression profiles under relevant physiological conditions and 2) test mutants in which a selected gene has been knocked out versus their wild type counterpart. cDNA probes for querying the array were derived from mRNA of cells grown in most instances in chemostats. For each condition tested competitive hybridizations were performed with Cy-Dye labeled cDNA probes. Post-hybridization the intensity of the two dyes for each gene (treatment vs. control) was measured by scanning the slide with a laser. The TM4 package (www.tigr.org/software/tm4) which consists of a suite of open-source programs developed at The Institute for Genomic Research was employed for microarray data analysis. Background corrected intensity values were normalized prior to examination for significant changes indicating up or down regulation of a gene.

An initial test of the array queried cells grown under nitrogen fixing conditions to a wild type control. The expression of at least ninety genes was determined to have changed significantly using the Significance Analysis of Microarrays (SAM) algorithm which incorporates gene-specific t-tests and false discovery rates to determine significant changes in expression. These included genes known to be vital to the nitrogen fixation process for example the up regulation of the genes responsible for nitrogenase, the key enzyme in nitrogen fixation. In addition, others genes not immediately predicted to be expressed under nitrogen limiting conditions were determined including the up regulation of several hypothetical genes and a sensor histidine kinase and response regulator.

Analysis of microarray data is also proving beneficial in corroborating predictions from annotation and analysis of whole genome data. Genome analysis predicted that *G. sulfurreducens* is not a strict anaerobe and is capable of using oxygen. A microarray analysis of *G. sulfurreducens* growth on 5% oxygen versus standard anaerobic conditions revealed up regulation of genes predicted to be involved in an oxidative metab-

olism including the high oxygen affinity oxidase, cytochrome d ubiquinol oxidase as well as ruberythrin which is capable of scavenging oxygen radicals.

In addition to examining gene expression profiles from individual physiological conditions and mutants, another facet of this project is to apply further data mining methods to the resulting array data across multiple experiments with the goal of elucidating functional roles and regulatory patterns in this organism. A variety of statistical and clustering techniques are currently being utilized. For instance, the software application, Expression Analysis Systematic Explorer (EASE), is being evaluated as a tool for determining biological themes present in significantly up and down regulated genes across multiple experiments. In one example, conditions tested included: growth with a chelated iron source as electron acceptor, attached growth on energy-harvesting electrodes and growth in the presence of 5% oxygen all versus the same standard control (growth in suspension under anaerobic conditions). EASE was used to elucidate statistically significant over representation of biological categories based on Gene Ontology (GO) assignments from the up regulated and down regulated gene lists. A modified Fisher exact probability test (EASE score) and correction via bootstrapping was used to determine significance ($p < 0.05$). The GO biological process of electron transport and the GO molecular function of transport activity were among the over represented assignments in the up regulated gene lists. These results confirm the importance of electron transport across diverse physiological conditions while suggesting the importance of other metabolic processes such as transporter activity in this organism.

An Analysis of Variance (ANOVA) of this same data set was used to look for genes with statistically significant changes in their gene expression profiles between the three experiments versus those that did not. The two resulting categories of gene expression data were then examined using clustering techniques. This analysis revealed a cluster of genes that based on their physical location and coordinate regulation describe a putative operon with genes for a c-type cytochrome, a C_4-dicarboxylate transporter and several genes of unknown function. The operon is down regulated under growth with a chelated iron source, up regulated when grown as a biofilm on graphite electrodes and expression does not vary greatly when grown in the presence of 5% oxygen suggesting that it may in part be repressed by the presence of iron and important in biofilm growth. Conversely, another cluster includes a group of transporters putatively related to heavy metal efflux and a transcriptional regulator from the mercuric reductase family of regulators all of which are up regulated in a similar manner across each of the three experiments.

Additional data mining techniques being evaluated include the use of template matching algorithms in which the expression of one or more genes can be used as a template and genes with expression profiles similar to the template can be identified between the template and genes in the data set. These matches may indicate genes related by function and/or regulation. In the three experiment data set the mean expression values of a family of heat shock protein genes were used as a pattern to look for genes regulated in a similar fashion. This technique was successful in matching other heat shock and chaperone genes with similar expression profiles as well as two periplasmic c-type cytochromes and several genes whose functions may be related to membrane structure suggesting that in addition to coordinate regulation these genes may collectively participate in the assembly or degradation of the functional c-type cytochromes. These findings reveal the power of microarray technology coupled with a variety of data mining techniques to suggest new functional roles and regulatory patterns in this organism.

38

In Silico Elucidation of Transcription Regulons and Prediction of Transcription Factor Binding Sites in *Geobacter* Species Using Comparative Genomics and Microarray Clustering

Julia Krushkal[1]* (jkrushka@utmem.edu), Bin Yan[1], Daniel Bond[2], Maddalena Coppi[2], Kelly Nevin[2], Cinthia Nunez[2], Regina O'Neil[2], Barbara Methé[3], and **Derek Lovley**[2]

*Presenting author

[1]University of Tennessee Health Science Center, Memphis, TN; [2]University of Massachusetts, Amherst, MA; and [3]The Institute for Genomic Research, Rockville, MD

Geobacter species are important for bioremediation of a variety of environments contaminated with metal, metalloid, and organic waste compounds, and their ability to harvest electricity also suggests them as a possible source of alternative fuel for the future. Therefore, we are developing a model of the physiological responses of *Geobacteraeae* to different environmental conditions in order to more rationally optimize bioremediation and energy-harvesting strategies. As part of this effort, we are using a computational approach that utilizes genome sequence information and whole genome expression data to elucidate the transcription regulatory circuitry of the *Geobacteraceae*.

We are employing a combination of complementary computational strategies to most efficiently predict operons, regulons, and transcription factor binding sites in *Geobacteraceae*. These computational methods can be divided into three categories, i.e. those that (1) are based on individual genome sequences of *Geobacter* species; (2) compare genome sequences from several closely related species of *Geobacteraceae,* and (3) use microarray clustering of *G. sulfurreducens* genes.

Single-genome analyses of the completed genome sequence of *G. sulfurreducens* and draft contig assemblies of *G. metallireducens* and *Desulfuromonas acetoxidans* provided whole genome predictions of operon organization. For example, we identified 1418 putative operons and transcription units in the *G. sulfurreducens* genome. As a first step in interpreting genome information obtained from the *Geobacteraceae* sequencing projects, we currently perform routine operon structure predictions with each new round of contig assembly of each genome (Table 1).

Table 1. An example of an NADH-quinone oxidoreductase operon predicted in the *D. acetoxidans* genome

Gene No. in the operon	Location (bp) in contig 548	Putative gene function
1	18892 –19248	NADH:ubiquinone oxidoreductase subunit 3 (chain A)
2	19239 –19748	NADH:ubiquinone oxidoreductase 20 kD subunit and related Fe-S oxidoreductases
3	19790 -20272	NADH:ubiquinone oxidoreductase 27 kD subunit
4	20306 –21496	NADH:ubiquinone oxidoreductase 49 kD subunit 7
5	21525 -22022	NADH:ubiquinone oxidoreductase 24 kD subunit
6	22067 -23848	NADH:ubiquinone oxidoreductase, NADH-binding (51 kD)
7	23882 -26362	Uncharacterized anaerobic dehydrogenase
8	26388 -27350	TPR-repeat-containing protein
9	27412 -28449	NADH:ubiquinone oxidoreductase subunit 1 (chain H)
10	28477 -28872	NADH:ubiquinone oxidoreductase 23 kD subunit (chain I)
11	28893 -29396	NADH:ubiquinone oxidoreductase subunit 6 (chain J)
12	29422 -29724	NADH:ubiquinone oxidoreductase subunit 11 or 4L (chain K)
13	29768 -31750	NADH:ubiquinone oxidoreductase subunit 5 (chain L)
14	31797 -33353	NADH:ubiquinone oxidoreductase subunit 4 (chain M)
15	33402 -34862	NADH:ubiquinone oxidoreductase subunit 2 (chain N)

We further analysed genome sequences of *G. sulfurreducens* and *G. metallireducens* by providing predictions of potential transcription regulatory elements using similarity searches to over 60 position-specific matrices of established transcription factor binding sites from other prokaryotes and by the neural network approach. As a result of these searches, we have developed databases of predicted transcription regulatory elements in each genome, along with software tools for querying these database in user-specified locations (Table 2).

Table 2. An example summary of the most significant positive predictions of transcription regulatory elements in the *G. sulfurreducens* genome based on whole genome similarity searches:

Transcription factor binding site	Number of highly significant hits in the *G. sulfurreducens* genome		Transcription factor binding site	Number of highly significant hits in the *G. sulfurreducens* genome
ArgR	2		metJ	4
CpxR	3		metR	12
Crp	58		narL	2
CytR	12		ompR	39
dnaA	57		rpoD15	547
FarR	62		rpoD16	446
Fis	123		rpoD17	1590
Fnr	2		rpoD18	244
FruR	2		rpoD19	422
Fur	4		rpoS17	227
GlpR	33		rpoS18	3
Hns	1396		soxS	41
Ihf	240		torR	1
Lrp	1117		tyrR	9
MalT	514			

Using **comparative genome analyses**, we verified our operon predictions in *Geobacteraceae* genomes by identifying clusters of genes conserved across three species from that family*: G. sulfurreducens, G. metallireducens*, and *D. acetoxidans.* Many of these genes participate in essential cell functions, e.g., DNA replication and protein biosynthesis. Interestingly, one of these conserved gene clusters involved genes related to flagellar proteins that are likely related to cell motility. We have developed and are maintaining a database of putative orthologs in these genomes. To date, the across-genome comparisons of gene clusters have allowed us to identify both conserved operons and operons unique to individual species of *Geobacter.* Using information from multiple genomes, we also searched for potential transcription regulatory elements by using the phylogenetic footprinting approach that identified conserved regions of noncoding DNA in different species of *Geobacteraceae*.

Further effective validation of operon and regulon predictions came from whole genome **microarray analyses** that involved hierarchical clustering of *G. sulfurreducens* genes based on their change in expression levels in *G. sulfurreducens* mutants as compared to the wild type. This approach allowed us to identify two groups of operons positively controlled by the fur regulator and one group negatively affected by this protein. Similarly, several groups of operons positively and negatively controlled by the RpoS regulator were identified. Microarray expression data are being used to predict transcription factor binding sites by identifying DNA elements conserved upstream of clusters of putative operons with similar expression patterns. A number of intriguing observations were made from these analyses. For example, both groups of operons positively controlled by fur contain fur binding sites and several other conserved motifs in their upstream noncoding regions, while the operons negatively controlled by fur do not seem to contain a fur box; instead they contain a motif highly similar to the lrp binding site. A group of operons under a strong positive control of RpoS contain in their upstream regions a -35/-10 box

along with other conserved motifs, suggesting that cooperative binding of transcription regulators affects the transcription of these operons.

The information obtained from the three computational strategies outlined here is being regularly compared and reconciled. This approach allows us to most efficiently identify putative transcription regulatory interactions among genes of *Geobacteraceae* and to identify groups of co-regulated genes and their putative DNA regulatory elements, as our first step toward the understanding of the complex network of regulatory interactions of *Geobacter.*

39

In Silico Modeling to Improve Uranium Bioremediation and Energy Harvesting by *Geobacter* species

R. Mahadevan[1], B. O. Palsson[1], C. H. Schilling[1], D. R. Bond[2], J. E. Butler[2], M. V. Coppi[2], A. Esteve-Nunez[2], and **D. R. Lovley**[2] (dlovley@microbio.umass.edu)

[1]Genomatica, Inc., San Diego, CA and [2]University of Massachusetts, Amherst, MA

Geobacter species are important organisms in the bioremediation of uranium-contaminated subsurface environments and for harvesting electricity from waste organic matter, but their metabolism is poorly understood. In order to better predict the response of *Geobacter* species under different environmental conditions and to further optimize bioremediation and energy harvesting applications, a genome-scale metabolic model of *Geobacter sulfurreducens* was developed using the constraints-based modeling approach. The metabolic model currently contains 523 reactions and 540 metabolites accounting for 583 genes (29% of the annotated genome).

The model has provided a number of new insights into the physiology of *G. sulfurreducens*. For example, one of the unsolved mysteries of anaerobic respiration in this organism is why growth yields with fumarate as the electron acceptor are 3-fold higher than during growth on Fe(III), despite the fact that Fe(III) has a higher mid-point potential than fumarate. The model has revealed the previously unsuspected importance of proton balance in the energetics of this organism and that more cytosolic protons are likely to be formed when Fe(III) is the electron acceptor. The energetic cost associated with the pumping of these extra protons to maintain the transmembrane gradient in cells growing on Fe(III) leads to a much lower energy yield than when fumarate serves as the electron acceptor. Thus, the model-based analysis has provided a likely explanation for the difference in biomass yields for growth with different electron acceptors. The current version of the model not only predicts these differences in growth yields, but also accurately predicts growth rates with Fe(III) or fumarate.

One of the most useful applications of the model has been to predict the phenotype of mutations generated for functional genomics studies. For example, a knock-out mutant that no longer produced succinate dehydrogenase could not grow with acetate as the electron donor and Fe(III) as the electron acceptor. However, the model made the non-intuitive prediction that this mutant would be able to grow better than the wild type on acetate and Fe(III) if fumarate was also provided. This prediction was experimentally verified. These types of predictions have the potential to

greatly accelerate functional analysis of genes and the understanding of the central metabolism of *G. sulfurreducens*.

Furthermore, *in silico* deletion studies can make laboratory mutational studies more efficient. For example, *in silico* deletion analysis revealed that deletion of genes associated with central metabolism led, in most cases, to either a lethal or a silent phenotype. Thus, this result has suggested that investing labor and time in making mutations in many of these genes may not be a fruitful line of investigation.

The model has also been helpful in elucidating the function of genes of unknown function and interpreting the results of microarray analysis of gene expression. For example, a knock-out mutation in a gene previously annotated as an Fe(III) reductase did not have the expected specific effect on Fe(III) reduction, but rather appeared to have a more general effect on metabolism. When the results of a microarray study comparing gene expression of this mutant with the wild type, were analyzed with the model, the gene expression changes were found to map closely with predicted changes in metabolism in an *in silico* mutation in NADPH dehydrogenase. This prediction of function from the model and other evidence has indicated that this gene encodes for a NADPH dehydrogenase, rather than an Fe(III) reductase, as previously proposed.

The model has also provided further insight into why *Geobacter* species predominate over other Fe(III)-reducing microorganisms, such as *Shewanella* and *Geothrix* species, in a diversity of subsurface environments. Previous studies have suggested that *Shewanella* and *Geothrix* species release extracellular electron-shuttling compounds in order to reduce Fe(III) whereas *Geobacter* species do not. Analysis of the energetic cost of producing an electron shuttle under conditions typically found in subsurface environments demonstrated that a microorganism, like *Geobacter* species, that did not need to produce a shuttle would grow 20-50% faster than an organism that produced an electron shuttle. This is a substantial difference that would provide *Geobacter* species with a significant competitive advantage.

The model is also helpful for predicting environmental manipulations that might stimulate the growth of *Geobacter* species, possibly accelerating bioremediation. For example, simulation studies demonstrated that there are a few amino acids, which if provided to *Geobacter*, would enhance its growth. These results are now being evaluated experimentally to determine if they represent a potential strategy to increase biomass yields.

These results demonstrate that this iterative modeling and experimentation approach to microbial physiology can rapidly accelerate discovery of gene function and provide important physiological and ecological insights. It is clear from these results that the *in silico* model of *G. sulfurreducens* has the potential to help guide the development of better strategies for the bioremediation of uranium and other contaminants as well as aid in the design of improved *Geobacter*-based fuel cells.

40

Continued Studies on Improved Methods of Visualizing Large Sequence Data Sets

George M. Garrity[1] (garrity@msu.edu), Timothy G. Lilburn[2] (Tlilburn@atcc.org), and Yuan Zheng[1] (zhangyu6@msu.edu)

[1]Michigan State University, East Lansing, MI and [2]American Type Culture Collection, Manassas, VA

We have continued our investigations into the use of graphical and analytical techniques drawn from the field of Exploratory Data Analysis to gain insight into the taxonomic relationships among prokaryotes, as currently defined by the 16S SSU rRNA. In our initial studies, we found that the dimensionality of extremely large sequence datasets ($n > 10^5$) could be reduced by methods such as Principal Components Analysis (PCA), allowing accurate projection of the data into 2D maps of the taxonomic space. In addition to revealing the overall topology of the taxonomic space, each strain could be accurately located within that space. While such plots were useful in delineating major groups, they proved less useful in resolving precise placement of individuals within some families and genera. Subsequently, we reported on the use of heatmaps, a form of colorized matrix, to directly visualize sequence similarity data. These plots readily revealed that many of the discrepancies detected by PCA could be attribute to a variety of taxonomic and sequence annotation errors. We also reported on the development of a self-organizing self-correcting classifier (SOSCC) that allowed for automatic detection and resolution of such errors; the SOSCC automatically optimizes the classification of the sequences, correctly positioning misplaced sequences based on their relationship to a set of validated nearest neighbors. Here we report on: (1) recent refinements in the SOSCC algorithm and porting of the algorithm to StatServer for deployment as an interactive web application, (2) the production a web-based taxonomic atlas of prokaryotes based on PCA plots and interactive heatmaps, (3) how this methodology has been applied to resolve a number of outstanding taxonomic anomalies, and (4) the development of a set of vetted sequences that can be used to improve the accuracy of identification of prokaryotes by 16S rRNA sequence analysis. Two spin-off projects applying this technology will also discussed: (1) integration of the application with the RDP to pipeline sequence data and (2) the use of interactive heatmaps as a graphical interface to access networked data resources.

41

PQuad for the Visualization of Mass Spectrometry Peptide Data

Bobbie-Jo Webb-Robertson[1] (Bobbie-Jo.Webb-Robertson@pnl.gov), Susan L. Havre[2], Deborah A. Payne[3], and Mudita Singhal[2]

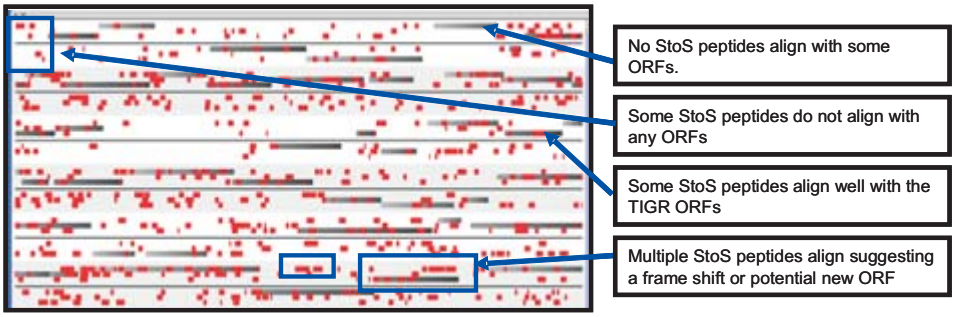
[1]Statistics & Quantitative Sciences, [2]Scientific Computing Environments, and [3]Information Analytics, Pacific Northwest National Laboratory, Richland, WA

Mass spectrometry has come forward as one of the most promising technologies for high-throughput proteomics. Thus, the development of supporting techniques and software tools for analyzing MS/MS data has been a high priority in recent years. The basic process clips a protein into peptides via proteolyic digestion prior to subjecting to the MS process. The resulting spectra are then subjected to a peptide identification tool. When the number of identified peptides becomes large, navigating and analyzing the dataset becomes a time-consuming and challenging task. The problem is exacerbated when the scientist attempts to integrate the results with other biological data or compare two or more sets of identified peptides, for instance, peptide sets collected under different experimental conditions. Information visualization and statistical methods are emerging as the technologies of choice in integrating and analyzing very large complex data. No tools comprising both these technologies currently exist to help scientists analyze MS/MS results. We have developed a proof-of-concept interactive visualization prototype, PQuad (Peptide Permutation and Protein Prediction), for the analysis of identified peptides in the context of an annotated DNA sequence. PQuad shows great promise in assisting in the evaluation and validation of both gene annotation and peptide identification software.

Currently, PQuad visualizes the identified peptides, the associated open reading frames (ORFs), and the actual nucleotide sequence of the DNA. The prototype provides three basic levels of resolution or views: summary, intermediate, and detail. The summary view is a miniaturized visualization of the ORFs and/or peptides overlaid on the complete DNA sequence; it provides a bird's eye view. The detail view shows a readable stream of both DNA nucleotide strands and the six possible amino acid reading frames; the location of the ORFs and peptides are indicated by highlighting the appropriate sections of the DNA strands and reading frames. The intermediate view is more flexible in both the resolution and the amount of information. Sequence letters are not provided in the intermediate view, but the user can view a section of the DNA that shows individual ORFs in the context of neighboring ORFs. Each view continuously reports the sequence indices, ORF names, and peptides under the cursor for easy exploration. Selecting an ORF in either the summary or intermediate view is propagated to all views so that the selected ORF is featured in the detail view; centered, highlighted, and surrounded by neighboring ORFs in the intermediate view; and highlighted in the summary view.

More recently, we have been working with a *Deinococcus radiodurans* spectral dataset that has been run through SEQUEST (a popular commercial peptide identification tool) with both TIGR and Stop-to-Stop (StoS) annotations. The difference in the number of identified peptides and ORFs is striking. The StoS identifications overlaid on the StoS annotation appears noisy, most likely due to a large number of false positives; many ORFs are observed with a single peptide hit. However, there are many places where groups of peptides cluster in apparent confirmation of the underlying ORF.

As a next step, we relaxed our rule that PQuad show the peptides against the ORF annotation that was used by SEQUEST to identify the peptides. Now PQuad can show the peptides (red) identified using the StoS annotation against the TIGR ORF annotation (gray boxes). The results of this combination are startling. It is easy to see where groups of StoS peptides align with TIGR ORFs. A few ORFs do not have associated peptides, but interestingly groups of StoS peptides cluster where there is no TIGR ORF. We believe that these peptides may suggest missed start positions, potential new genes, or gene modifications not included in the TIGR annotation.



42

Computational Framework for Microbial Cell Simulations

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

[1]Pacific Northwest National Laboratory, Richland, WA and [2]University of Texas Medical School, Houston, TX

Development of integrated sets of computational tools is needed to achieve the level of sophistication necessary to bridge experimental and computational biology studies. Because of the complexity of the biological data associated with the cellular processes, use of mathematical and computational methods are needed to decipher the information hidden in the experimental results and to design new experiments. As part of this project, we have been developing a wide range of prototype computational biology and bioinformatics analysis tools, and new algorithms and methods. New tools are employed to investigate the flux and regulation of fundamental energy and material pathways in *Rhodobacter sphaeroides*. The prototype components are designed in such a way that, when combined later, they will form the backbone of a comprehensive microbial cell simulation environment.

Our recent efforts to develop a computational framework have concentrated on the following research areas:

Gene regulatory networks: We have developed a new probabilistic algorithm to model the stationary properties of the gene regulatory networks and our new algorithm was implemented in the object oriented stochastic simulation software NWGene. We applied the new algorithm to simulate the expression patterns in a library of synthetically engineered gene regulatory networks. The agreement between the model predictions and the experimental data was very good.

Stochastic kinetic simulations: We have further improved the computational efficiency of the NWKsim program, a kinetic simulation package that uses stochastic Gillespie algorithm and its variants. We used the NWKsim program to investigate the cell receptor signaling networks using kinetic models.

Imaging of bacterial cells and image recontruction: We have obtained electron tomography images of *R. sphaeroides* using the TEM imaging facility at UCSD. Utilizing our image reconstructed software NWGrid, we have computationally reconstructed a 3-D geometry of *R. sphaeroides'* surface features from the obtained series of tilted digital images.

Mesh grid based simulation framework: We are using the VMCS (Virtual Microbial Cell Simulator), which is based on the biological version of NWGrid/NWPhys (http://www.emsl.pnl.gov:2080/nwgrid), to simulate spatial and temporal growth of microbial cell communities. The model includes explicit representations of individual microbial cells, derived from TEM image data or computational geometry. The evolution of the model allows for the generation of communities that can take the form of biofilms or free floating flocs. The VMCS model imports reaction/diffusion models and can include environmental conditions such as flow velocity, shear flows, and structures. As a test application of the software, we are simulating the biofilm growth in a two-organism syntrophic bacterial system.

Genome comparison data mining: Genome comparison data mining detects larger patterns useful in understanding complex biological processes from large quantities of sequence data across many species. Our Similarity Box software provides a sensitive and accurate method for extracting several important types of genome comparison results, including conserved gene neighborhood relationships, which are informative for protein function and binding partners. We have now incorporated a high-throughput version of this approach to gene neighbor analysis in a HERBE database implementation designed to support *R. sphaeroides* genome annotation. A user can enter a single *R. sphaeroides* protein identifier and receive a useful view of all relevant conserved relationships. Using the new implementation, for example, we found that the *R. sphaeroides* RpoH1 protein belongs to a cluster of heat shock factors with a conserved association with pseudouridine synthases, in contrast to the *E. coli* RpoH which is linked to the FtsYEX proteins. This strategy will be made available to the *R. sphaeroides* annotation community.

Determination of regulatory mechanisms for the photosynthesis genes of *R. sphaeroides*: Although most of the regulators of the photosynthesis genes of *R. sphaeroides* have been determined using biochemical methods, detailed understanding of the regulatory mechanisms is still lacking. We are using the recent genome (DOE sponsored JGI) and microarray (UT-Houston) data to investigate the regulators of the photosynthesis genes of *R. sphaeroides*. Using a combination of clustering analysis and DNA motif finding methods, we have investigated the DNA recognition motifs of the known regulators. We were able to confirm the binding motifs of the regulators FNR and PpsR, and we have derived a statistical distribution of the binding motif of another regulator PrrA. We have also developed a new approach to search for motifs in DNA sequences. Our approach, which is computationally intensive, combines techniques from combinatorial and numerical optimization, and was implemented with a parallel genetic algorithm.

43

Optimization Modules for SBW and BioSPICE

Vijay Chickarmane, **Herbert M. Sauro** (hsauro@kgi.edu), and Cameron Wellock
Keck Graduate Institute, Claremont, CA

Parameter estimation and model validation are essential components to model building. As part of the DARPA BioSPICE project, we have developed a series of optimization modules which enable experimentalists to fit time series data to ordinary differential equation (ODE) based models. Given a model and a set of experimental data, the optimization modules compute estimates for the model parameters, the sums of squares of the final fit and standard errors on the certainly of the fitted parameter values.

The modules are written in Matlab and employ a new Systems Biology Workbench (SBW)/Matlab interface which makes integration of Matlab scripts into SBW extremely easy and enables us to leverage existing SBW tools such as model designers and simulators. The modules themselves employ a number of novel approaches to optimization, some of which we believe are suitable for optimizing large systems. The SBW integration permits the modules to be automatically used by the BioSPICE Dashboard interface and thus to appear as building blocks in a Dashboard workflow diagram. The use of SBW also permits developers to write additional modules and have them automatically integrated into the BioSPICE/SBW with very little effort. Note that the operation of the modules does not require Matlab to be installed on the client machine.

In addition to the modules themselves, we also provide an Optimization Controller GUI which enables users to easily employ these modules in their research. The controller permits different optimization methods to be applied either individually or in succession. During the optimization a real-time graphical display is generated that enables one to judge the effectiveness and progress of the optimization. Optimizations may be stopped and started, and different methods applied during the optimization. Users can also graphically compare the fits with the experimental data.

Due to integration into SBW, models can be developed under a variety of tools (eg JDesigner, CytoScape, CellDesiginer) via SBML. Simulation engines such as Jarnac or Dizzy can be exploited by the optimization modules to compute the solutions to the ODEs which leads to significantly improved performance. Due to the integration into SBW/BioSPICE additional control of the optimization procedures is available via Python and Perl scripts. This option provides great flexibility, for example a user can implement Monte Carlo fitting for those systems where the non-linearities in the model do not permit accurate estimates for the standard errors. In the first release the following optimization modules will be made available:

- Levenberg-Marquardt
- Nelder & Mead Simplex
- Simulated Annealing/Simplex Hybrid
- Genetic Algorithm
- Genetic Algorithm/Simplex Hybrid

The software will be released on our web site (www.sys-bio.org) by the time of the 2004 GTL meeting.

44

The Docking Mesh Evaluator

Roummel Marcia[1] (marcia@math.wisc.edu), Susan D. Lindsey[2] (lindsey@sdsc.edu), J. Ben Rosen[3] (jbrosen@ucsd.edu), and **Julie C. Mitchell**[1] (mitchell@math.wisc.edu)

[1]Departments of Mathematics and Biochemistry, University of Wisconsin, Madison, WI; [2]San Diego Supercomputer Center, University of California, San Diego, CA; and [3]Department of Computer Science and Engineering, University of California, San Diego, CA

Introduction

The Docking Mesh Evaluator (DoME) is a software for predicting a bound protein-ligand docking configuration by determining the global minimum of a potential energy function. Our present energy model is based on solvent effects defined implicitly using the Poisson-Boltzmann equation, as well as a pairwise Lennard-Jones term.

Description

Our approach consists of two phases. The first involves scanning the energy landscape for favorable configurations. This phase can be done once as a preprocessing step and need not be done again. The second phase involves the iterative underestimation of successive collections of local minima with convex quadratic functions, using the configurations from the first phase as initial seed points for optimization. The minima of the underestimators are then used as predicted values for the global minima. Both serial and parallel versions of this "coupled" optimization have been successfully implemented. Preliminary results are reported in [2].

Currently, our research is focused on optimizing parameters in the energy function, in order to obtain the best accuracy in predicting known docking configurations. In particular, we consider the benchmarking set of Chen et al. [1] for testing protein-protein docking algorithms. Of the 59 test cases it contains, 22 are enzyme-inhibitor complexes, 19 are antibody-antigen complexes, 11 are various diverse complexes, and 7 are difficult test cases whose solutions have significant conformational changes. These optimized parameters are expected to yield realistic results for biological problems whose solutions are unknown.

Flexibility in the protein-ligand model is being implemented using a hybrid of global optimization and rotamer search. Near the surface interface, subtle side-chain rearrangements are often necessary to model induced fit between the receptor and the ligand. These rearrangements can be modeled using candidate residue conformations, called rotamers. Using this approach, the protein backbone is held fixed while residues are allowed to take on various configurations. Such pseudo-flexibility is a more viable alternative to full backbone and side-chain flexibility, which requires inordinately many free variables, thus making the computational cost prohibitively expensive. Local shape complementarity analysis performed using the Fast Atomic Density Evaluator [3] will provide added efficiency by highlighting regions in which shape mismatches occur.

References

1.

R. Chen, J. Mintseris, J. Janin, and Z. Weng, "A protein-protein docking benchmark," *Prot. Struct. Fun. Gen.*, **52**, pp. 88–91, 2003.

2. R. F. Marcia, J. C. Mitchell, and J. B. Rosen, "Iterative convex quadratic approximation for global optimization in protein docking," *Comput. Optim. Appl.*, Submitted, 2003.

3. J. C. Mitchell, R. Kerr, and L. F. Ten Eyck, "Rapid atomic density measures for molecular shape characterization," *J. Mol. Graph. Model.*, **19**(3), pp. 324–329, 2001.

45

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, St Paul, MN

Microbial metabolism is vast and much remains to be catalogued and characterized. Characterizing this metabolism is a major task of microbial functional genomics. Over time, these data will impart much greater predictive power onto microbial science. The research conducted on this project seeks to better assemble existing metabolic data, discover new microbial metabolism, and predict microbial metabolic pathways for compounds not yet in the databases.

One goal of the project, compilation of information relevant to the metallic and metalloid elements that comprise half of the periodic table, has been completed. The web-based University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD) has been expanded to include information on microbiological interactions with 77 chemical elements (1). This is part of a comprehensive study of how microbes interact with all of the chemical elements (2). For each element, a webpage has been created with annotation on the major microbial interactions with that element, links to Medline, and access to further UM-BBD information. The project has added hundreds of new linkages to UM-BBD compound pages (3). For example, the mercury element page has 4 links, arsenic has 9 links, and chlorine has 145 links to UM-BBD compounds, respectively.

Another important goal of the current project is to discover new metabolism and functionally analyze the novel microbial enzymes and genes involved (4). A review of the natural product literature has revealed that on the order of one hundred chemical functional groups are produced by biological systems (5). Yet, only about fifty functional groups have been studied with respect to metabolism. Metabolism must exist for the remaining 50 chemical groups. This lack of information represents a knowledge gap contributing to the problem of incomplete microbial genome sequence annotation. In the current project, we are uncovering metabolism of chemical functional groups that have previously not been studied. To date, we have discovered new metabolism relevant to bismuth compounds, boronic acids (6), azetidine ring compounds, and novel organonitrogen compounds.

A third goal of the project has been to develop a computer software that predicts microbial metabolism using the UM-BBD as a knowledge base (7). The user of the software gets to see one or more plausible biodegradation pathways for the com-

pound they have entered into the system. The metabolism prediction software is based on rules that broadly describe microbial reactions such they can be applied to new compounds. At present, there are over 250 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 has been validated by comparison against: (i) known biodegradation reactions, (ii) expert predictions, and (iii) microbial growth studies. The system is freely available on the web (8). The system will be expanded with input from our Scientific Advisory Board and the broader scientific community.

References

1. UM-BBD Biochemical Periodic Tables: http://umbbd.ahc.umn.edu/periodic/

2. Wackett, L.P. A.G. Dodge, and L.B.M. Ellis. (2004) Microbial genomics and the periodic table. *Appl. Environ. Microbiol.* (in press).

3. Ellis, L.B.M., B.K. Hou, W. Kang and L.P. Wackett (2003) The University of Minnesota Biocatalysis/Biodegradation Database: Post genomic data mining. *Nucl. Acids Res.* **31**:262-265.

4. Wackett, L.P. (2002) Expanding the map of microbial metabolism. *Environ. Microbiol.* **4**: 12-13.

5. Wackett, L.P. and C. Douglas Hershberger (2001) *Biocatalysis and Biodegradation: Microbial Transformation of Organic Compounds*. American Society for Microbiology Press.

6. Negrete-Raymond, A.C., B. Weder, and L.P. Wackett (2003) Catabolism of arylboronic acids by *Arthrobacter nicotinovorans* strain PBA. *Appl. Environ. Microbiol.* **69**:4263-4267.

7. Hou, B.K., L.P. Wackett, and L.B.M. Ellis (2003) Microbial pathway prediction: A functional group approach. *J. Chem. Inf. Comp. Sci.* **43:**1051-1057.

8. UM-BBD Pathway Prediction System: http://umbbd.ahc.umn.edu/predict/

46

Comparative Genomics Approaches to Elucidate Transcription Regulatory Networks

Lee Ann McCue* (mccue@wadsworth.org), Thomas M. Smith, William Thompson, C. Steven Carmack, and **Charles E. Lawrence**

*Presenting author

The Wadsworth Center, New York State Department of Health, Albany, NY

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. This approach has two components: motif identification via phylogenetic footprinting, and regulon identification via the clustering of motifs. The phylogenetic footprinting step requires the genome sequences of several closely related species, and employs an extended Gibbs sampling algorithm to analyze orthologous promoter data to identify individual transcription factor binding sites and the associated motif model of

common binding patterns. The accuracy of these predictions has been evaluated by comparison with sets of sites reported for 166 genes in *Escherichia coli*, revealing that 75% of predicted sites overlap the experimentally verified sites by 10 bp or more. We have also developed a novel Bayesian clustering algorithm to predict regulons via clustering of the motifs identified in the footprinting step. Again, we validated this technique by comparison with reported regulons in *E. coli*. This inference of regulons utilizes only genome sequence information and is thus complimentary to and confirmative of gene expression data generated by microarray experiments. Here we describe preliminary results of our applications of these technologies to the *Synechocystis* PCC6803 genome.

Project ID: DE-FG02-01ER63204

47

Elucidating and Evaluating Patterns of Lateral Gene Transfer in Prokaryotic Genomes: Phylogenomic Analyses using GeneMarkS Gene Predictions

John Besemer[1], **Mark Borodovsky**[1] (mark.borodovsky@biology.gatech.edu), and John M. Logsdon, Jr.[2]

[1]Schools of Biology & Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA and [2]Department of Biological Sciences, University of Iowa, Iowa City, IA

Is the extent of lateral gene transfer (LGT) in prokaryotic genomes so large as to preclude using phylogenetic methods to elucidate evolutionary relationships among major bacterial lineages? The evolutionary mode of many ribosomal components has apparently been vertical since phylogenetic trees of these genes are in considerable agreement. Nonetheless, phylogenetic trees derived from many genes are either non-resolvable or incongruent (*e.g.* with a presumed species tree). With a general goal of determining the utility of phylogenetic methods to understand prokaryotic evolution, and a specific goal of understanding the utility of predicted 'Atypical' genes as surrogates for laterally transferred genes, we have begun a phylogenetic comparative study of GeneMarkS-predicted genes in 121 selected prokaryotic genomes, including many from the DOE Joint Genome Institute. Each of the GeneMarkS-predicted genes (both typical and atypical classes) from six cyanobacterial genomes was subjected to an iterative BLASTP procedure designed to find complete homolog sets from a collection of 121 complete genomes. While 296 gene families that contained at least one atypical cyanobacterial member resulted, only 161 contained enough taxa to construct rooted phylogenetic trees. The trees were classified into three major groups: i) trees where the cyanobacteria formed a single clade (including the trivial subset which contained cyanobacterial genes exclusively), ii) trees where the cyanobacteria were split, and iii) trees with a single cyanobacterial taxon. In (i) cases, the focus of the search for potential LGT was among the cyanobacteria themselves. In the other (ii & iii) cases, potential transfers into and out of the cyanobacterial lineage were investigated. To extend our analysis to the larger question of LGT among all prokaryotes, we needed to derive a more reliable reference topology. Incongruities in trees built for particular genes compared to this reference tree are potentially indicative of LGT. We have been developing a reference tree from concatenated protein alignments built from groups of genes empirically selected based on their presence or absence in our set of 121 complete genomes (*e.g.* their phylogenetic distributions). We are experimenting

with several different criteria for selecting the genes and trimming the alignments, as well as testing two methods for phylogenetic tree construction: maximum likelihood distance (from TREE-PUZZLE) and Bayesian likelihood (from MrBayes). Our current reference topology has been generated from 37 different genes and totals more than 5000 amino-acid residues in length. With our ultimate goal to estimate the extent and pattern of LGT among all prokaryotes, we have randomly selected 3000 'typical' predicted genes and 5000 'atypical' predicted genes from a representative set of 26 genomes. Analysis of these sets is being used to determine if atypical codon usage is a reliable predictor of LGT as detected by the extent of phylogenetic incongruities with respect to the class of typical genes. The methods being developed herein are easily scalable to allow the inclusion of newly sequenced genomes into the analysis and allow the adjustment of important parameters (such as BLAST E-values). In addition to testing the validity of using atypical genes as surrogates for laterally transferred genes, the results of these analyses will provide solid, phylogenetically-based estimates for the rates of LGT in prokaryotic genomes.

48

Cell Modeling and the Biogeochemical Challenge

P. J. Ortoleva (ortoleva@indiana.edu)

Center for Cell and Virus Theory, Indiana University, Bloomington, IN

As the Genomics:GTL project matures, at CCVT we are completing our cell models Karyote® and CellX® and are planning the multiple space-time extensions needed to use them in environmental analysis. Our two cell models, a virus-intracellular feature model, and future perspective are as follows.

Karyote® is a cell model that accounts for transcription and translation (by step-by-step polymerization), metabolics and molecular exchange between organelles and cytoplasm or with the surroundings. The reaction-transport equations are solved using multiple timescale mathematical and computational techniques. It has interfaces that allow for the building of a compartmented eukaryotic cell or a simple or composite bacterium. Modules also have been developed for creating multi-cellular systems from models of single cell types, or for viewing results or the network of processes accounted for. The Karyote® system includes a database of cell properties, pathways and kinetic parameters, and procedures for model calibration using raw (e.g. NMR, mass spectral, microarray, microscopy) data.

CellX® is a finite element simulator that simultaneously solves reaction-transport equations on fibriles (1-D), membranes (2-D) and bulk medium (3-D), and the exchange among them through boundary conditions. CellX® has all the features of Karyote® and in addition accounts for gradients within each compartment (notably of proteins and other slowly migrating species)

VirusX® is designed to model the structure or function of a virus or other subcellular feature. It accounts for the supra-million atom structural detail of the virus or other object. VirusX® solves the molecular mechanics problem using space-warping and tree-code methods. We are developing mixed mesoscopic models wherein the capsid or other viral features can be described by continuous variables or wherein the dynamics of focus macromolecules is simulated using efficient computational techniques. The electrolyte host medium is treated using our continuum

position-orientation density description, a nonlinear, nonlocal dielectric model and associated free energy functional.

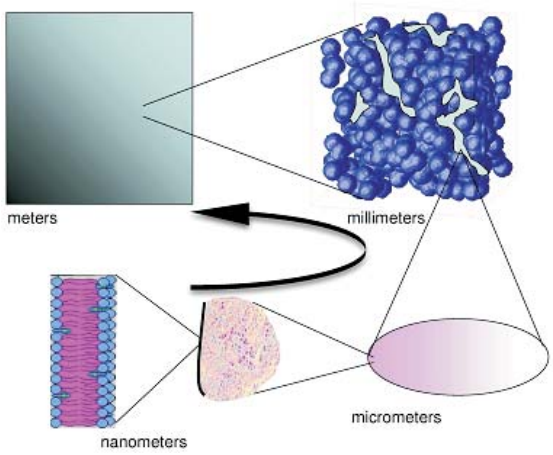
For computational biology to make a major contribution to microbial research, several grand challenges must be addressed. Greater predictive capability will only follow when reaction-transport models account for metabolic, proteomic, genomic and structure-forming (i.e. biomic) processes. Models accounting for this biomic process network will allow biologists to understand all the couplings among these processes and to simulate the life cycle of a microbe. This challenge is being addressed by our Karyote® model. The architecture of a microbe interior must be understood as a multi-dimensional system; this is being addressed via our CellX® simulator as described above. The impressive intracellular architecture and related functioning is strongly influenced by molecular-scale physics (e.g. fibrils along which molecules are trafficked between specific origins and destinations). Thus a fuller understanding of microbial behavior must involve the integration of molecular- and mesoscopic-scale physics and chemistry. In short, microbial models must be cast in the language of nanotechnology and mesoscopic theory, integrating atomistic to macroscopic variables (e.g. molecular structure to metabolite concentrations). In this way, microbial models will be cast in terms of the three scales (molecular, mesoscopic and macroscopic) at which biologists conceptualize and integrate their thoughts. This challenge is being addressed by our all-atom/mesoscopic simulator, VirusX®.

It must be admitted that for at least the next decade models will be incomplete and thus, even for well-understood processes, the rate and thermodynamic parameters are known only with great uncertainty or not at all. Even rate parameters determined from cell extracts may be far from those in the intracellular environment. Thus, procedures are needed that address calibration even for incomplete models. The data available (e.g. NMR, mass spectra) will only be indirectly related to the model parameters. The two endeavors (i.e. comprehensive biomic cell modeling and experimental data acquisition/interpretation) should be unified into one procedure to automate the building of models and the design of experiments to minimize (and assess) the uncertainties in both. Approaches must be developed that integrate the more complete biomic models into multi-cellular/sediment composite media systems. Effective models must account for the network of biomic intracellular processes that cannot be accounted for in available lumped models. Fundamental approaches starting from pore-scale detailed descriptions that are then rigorously upscaled to the field are needed to attain predictability as suggested in the figure. Mathematical/computational approaches involving homogenization theory should be applied to account for flow, diffusion, aqueous and mineral reactions, and the exchange with the evolving microbial colonies. Research should not be based on straightforward extension of simple lumped models, as they cannot easily be extrapolated to other sites using a calibration at a given site.

In summary, we must automate the integration of experimental and modeling technologies, develop a multi-scale platform to facilitate creative life sciences thinking, cast the approach in a manner that addresses the inherent uncertainties in experimental and computational life sciences, and simultaneously utilize the rapidly growing databases of genomic, proteomic, metabolic and structural information. For example, the genomic and proteomic information should be integrated into a meta-

bolic model via a detailed kinetic model of the polymerization of mRNA and proteins constituting transcription, translation and post-translational processing.

Schematic depiction of the multiple levels at which understanding of microbial systems is needed to attain quantitative predictability. Advances in the accuracy of the models at each level and methods for upscaling them are needed.



49

Rapid Reverse-Engineering of Genetic Networks via Systematic Transcriptional Perturbations

J. J. Collins (jcollins@bu.edu), T. S. Gardner, and C. R. Cantor

Department of Biomedical Engineering, Boston University, Boston, MA

The collection and assembly of large-scale genetic data into comprehensive databases is often regarded as the necessary first step in the elucidation of genetic network structure and function. However, it is not obvious if or how the disparate and partial data populating such databases can be assembled into unambiguous and predictive models of genetic networks. To address this problem, we have developed a reverse-engineering method that enables rapid construction of a first-order quantitative model of a gene regulatory network using no prior information on the network structure or function. The method, called Network Identification by multiple Regression (NIR), uses a series of steady-state transcriptional perturbations, coupled with RNA, protein, or metabolite activity measurements and multiple linear regression, to construct the model. In a pilot study, we successfully applied the NIR method to reverse engineer a 9-gene subnetwork in *E. coli* (Gardner et al., Science 301: 102, 2003).

Computational testing and our pilot *E. coli* study suggest that the NIR method can be applied on a large scale. In this project, we plan to extend the method to larger prokaryotic networks. Specifically, we plan to apply the method to *Shewanella oneidensis* electron-transport networks relevant to bioremediation. We will use *Shewanella* microarrays developed by DOE researchers to perform large-scale RNA profiling for our experiments. Our efforts on *Shewanella* will be designed to build

on and to support the existing experimental and computational efforts of Genomics:GTL (GTL) researchers.

The mapping and modeling of genetic networks in prokaryotes, a central objective of the GTL project, will provide the foundations for a variety of applications that advance the DOE mission needs in energy and the environment. Our method could significantly improve the efficiency of existing efforts of GTL researchers to map and model genetic networks in microbes.

50

Computational Hypothesis Testing: Integrating Heterogeneous Data and Large-Scale Simulation to Generate Pathway Hypotheses

Mike Shuler (info@gnsbiotech.com)

Gene Network Sciences, Ithaca, NY

Most prokaryotes of interest to DOE are poorly understood. Even when full genomic sequences are available, the function of only a small number of gene products are clear. The critical question is how to best infer the most probable network architectures in cells that are poorly characterized. The project goal is to create a computational hypothesis testing (CHT) framework that combines large-scale dynamical simulation, a database of bioinformatics-derived probable interactions, and numerical parallel architecture data-fitting routines to explore many "what if?" hypotheses about the functions of genes and proteins within pathways and their downstream effects on molecular concentration profiles and corresponding phenotypes. From this framework we expect to infer signal transduction pathways and gene expression networks in prokaryotes. The focus of this proposal is the:

1. Extension of accurate dynamical simulation methods to genome-size scales (i.e., 1000s).

2. Normalization of confidence levels for a wide variety of bioinformatics algorithms for extraction of a database of probable interactions.

3. Extension of numerical data-fitting techniques to large multi-scale cell simulations and exploitation of biological network properties for more efficient use of computational resources.

4. Integration of 1., 2., and 3. to derive an ensemble of hypothetical network structures and their corresponding molecular concentration profiles and phenotypic outcomes.

In order to create, refine and validate such a method for application to organisms of DOE interest where little functional data is available, our project will address a number of issues.

- Given the enormous cost in both time and money to collect genome wide data sets, it is important to determine what types of data and what quantities and qualities of data are necessary to lead to an inference of pathway circuitry at a given confidence level. Our proposed CHT platform would accomplish this determination through the integration of varying amounts of heterogeneous data for a

non-DOE model bacterial organism where much of the functional biochemical circuitry is known.

- Methods currently exist to extract static biochemical circuitry through the application of bioinformatics algorithms to whole genome sequences. This static circuitry does not directly link a particular molecular circuitry to the corresponding molecular concentration profiles and corresponding phenotypes. Our dynamic CHT is the tool that would quickly and inexpensively (in)validate the enumerable number of predicted network architectures that arise from the application of bioinformatics algorithms to whole genomes.
- As a more detailed functional understanding of DOE microbes is attempted, experimental efforts could greatly be accelerated if it were possible to investigate the multitude of hypotheses faster and cheaper than can be accomplished by experiment alone. We propose CHT as a method to investigate enumerable hypotheses on the computer with the goal of eliminating many improbable hypotheses and suggesting a more focused set of experiments.

CHT will be tested, validated, and applied to three different systems in decreasing order of completeness and transparency and ability to validate. The first is a synthetic network system that is created within our computer simulation framework (where by definition 100% of the circuitry and molecular functions are known). The second system is *E. coli K-12* (where 60–70% of the molecular functions are known). The third system and the one that is ultimately of direct interest to the DOE's objective is *Shewanella oneidensis* (where less than 10% of the circuitry and molecular functions are known).

51

Bacterial Annotation Tools

Owen White (owhite@tigr.org)

The Institute for Genomic Research, Rockville, MD

Manatee (MANual Annotation Tool, Etc Etc) is a graphical user interface designed to manage data in a common database that allows multiple users to simultaneously operate on that information. Production annotation teams at TIGR, as well as outside collaborators, have been using this system for the past year to obtain essential annotation information in a user-friendly way. The system supports making functional assignments using search results, paralogous families, and annotation suggestions generated from automated analysis. It also produces summaries that report the progress of each annotation project. Manatee runs using a web browser interface and easily allows installation of additional web scripts to facilitate frequent and rapid improvements. The Manatee suite contains documentation for installation and general use by scientists, and also contains complete documentation of the code libraries that can be modified by other software developers. The complete code base is available as open source and may be installed locally on Unix computer systems. Classes are now offered quarterly at TIGR giving instruction on the use of this software and on the "best practices" that have been formalized to generate uniform annotation.

Because of the relative ease of adding to the Manatee system, we are using this architecture as the basis of the Sybil software, a system that will allow scientists to discover, evaluate, and summarize intra-species variation. Sybil data storage is an

open source, modular relational schema called Chado that has been populated with data from closely related species. The Sybil API which makes calls to the database is operational, and many prototypic interfaces that display complex comparative data have been implemented. The system will be used for comparative analysis in two broad areas: 1) the interface will be incorporated into a web resource and used on-line by scientists interrogating data from TIGR resident in our database, and 2) users will be able to install the Sybil system on their local computers to perform custom analyses of their data. We anticipate a release of Sybil sometime before the end of year.

We have also developed a workflow system that performs routine operations required for most annotation pipelines. This system is based on simple configuration files and workflow templates. The configuration file is human-readable and contains the definitions used to generate a particular executable instance of a workflow, such as the Blast or HMMer programs. The workflow template is an XML document and describes the graph defining the overall pipeline. To begin the pipeline, the workflow execution engine executes an instance of the template; this "workflow instance" contains the complete description of the pipeline that has been invoked, and will be continuously updated with the status of the pipeline. Placing the status of completion in the workflow instance document allows for monitoring the process, and supports resuming fail or aborted workflows from the appropriate starting point. This engine is capable of handling a variety of complex workflows that may contain a combination of parallel and sequential processes and executes jobs in a distributed or grid computing environments.

The above projects have been developed in conjunction with DOE support for the Comprehensive Microbial Resource and NSF funding.

52

RELIC - A Bioinformatics Server for Combinatorial Peptide Analysis and Identification of Protein-Ligand Interaction Sites

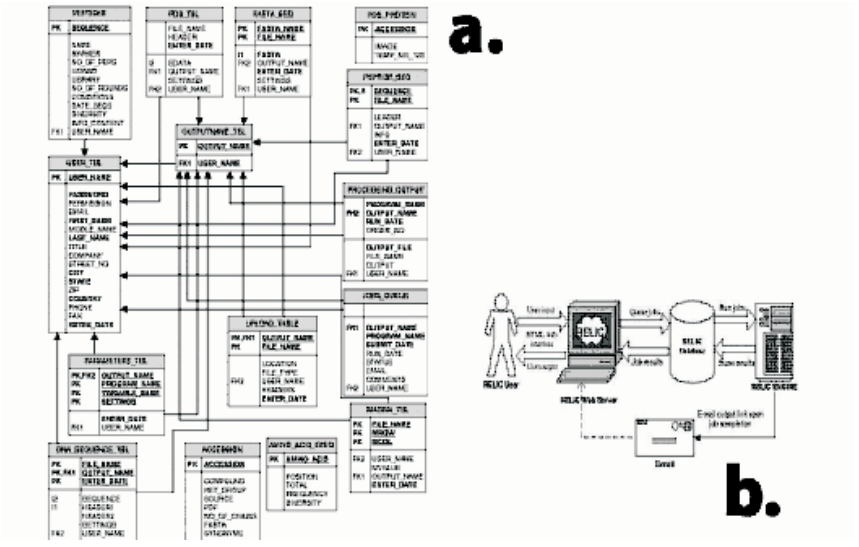
Suneeta Mandava*, Lee Makowski, Satish Devarapalli, Joseph Uzubell, and **Diane J. Rodi** (drodi@anl.gov)

*Presenting author

Biosciences Division, Argonne National Laboratory, Argonne, IL

The need for high throughput bioinformatic methods to characterize gene function is being driven by the generation of sequences at a rate far beyond our ability to carry out experimental functional analyses. In spite of the large number of analytical tools currently available, typically about 40% of predicted open reading frames remain functionally uncharacterized. An important clue to open reading frame function is the identification of binding partners. Phage display technology is a widely used tool for identifying either protein or small molecule binding partners. This project seeks to apply a novel approach to genome-wide identification of small molecule binding proteins. Preliminary results from our group has demonstrated that the similarity between the sequence of a protein and the sequences of affinity-selected, phage-displayed peptides can be predictive for protein binding to a small molecule ligand. Affinity-selected peptides provide information analogous to that of a consensus-binding sequence, and can be used in an analogous fashion to identify ligand binding sites.

In this project, libraries of phage-displayed peptides have been screened for affinity to the metabolites ATP and glucose, as well as other small molecule ligands. The sequences of affinity-selected peptides were determined and used as the basis of genome-wide analyses to identify proteins that have a high probability of binding to the screened ligands. The best set of affinity-selected peptides as validated through comparison with well-characterized proteins was used for genome-wide annotation of the *E. coli* genome as an initial test genome with a high percentage of functional annotation. During the course of this work, we have developed a suite of computational tools for the analysis of peptide populations and made them accessible by integrating fifteen software programs for the analysis of combinatorial peptide sequences into the REceptor LIgand Contacts (RELIC) relational database and web-server. These programs have been developed for the analysis of statistical properties of peptide populations; identification of weak consensus sequences within these populations; and the comparison of these peptide sequences to those of naturally occurring proteins. RELIC is particularly suited to the analysis of peptide populations affinity selected with a small molecule ligand such as a drug or metabolite. The order of the programs and their specific functionalities is specifically designed to aid a researcher in the combinatorial peptide field from the early stages of raw data acquisition to the final stage of protein epitope mapping. The flow of data-processing software starts with sequence translation programs, followed by physicochemical property mapping, sequence bias identification algorithms, and finally peptide/protein similarity mapping both within and in the absence of three-dimensional coordinates. In order to seamlessly integrate that biological data, RELIC is based on an object-oriented design using a relational database management system. For this particular project, the ORACLE 9i (Release 9.2) database system was chosen to store experimental data and the relevant genomic/structure information as it provides a wide array of database drivers for various programming languages (both for thin and thick clients). The figure at the lower left below (a.) is a diagram depicting the logical and relational model of the database by displaying all tables and intra-table relationships. The figure at the right (b.) shows a schematic of how users interact with the RELIC hardware. A RELIC user submits data for pro-



cessing via a web interface. The user input and job information is stored in a RELIC database. A job processing service periodically checks for pending jobs and processes them using the scientific algorithms developed in FORTRAN, using

COM+ interfaces. The user is sent an email upon completion of the job with a link to the output. Within this functional context, the ability to identify potential small molecule binding proteins using combinatorial peptide screening will accelerate as more ligands are screened and more genome sequences become available. The broader impact of this work is the addition of a novel means of analyzing peptide populations to the phage display community.

53

On Truth, Pathways and Interactions

Andrey Rzhetsky (ar345@columbia.edu)

Department of Biomedical Informatics and Columbia Genome Center, Columbia University, New York, NY

I will give an overview of our effort to automatically extract pathway information from a large number of full-text research articles (GeneWays system), automatically curate the extracted information, and to combine the literature-derived information with sequence and experimental (such as yeast two-hybrid) data using a probabilistic approach.

54

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

Susan M. Barns* (sbarns@lanl.gov), Elizabeth C. Cain, Leslie E. Sommerville, and **Cheryl R. Kuske** (kuske@lanl.gov)

*Presenting author

Los Alamos National Laboratory, Los Alamos, NM

The **overall goal** of this project is to identify novel, previously uncultured groups of bacteria that may play important roles in bacterial community function in contaminated sediments and soils, and to obtain genomic DNA of members of these bacterial groups for genome analysis. Our studies have focused on non-cultured members of the *Acidobacteria* 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). Their functions in soils and sediments are unknown.

The **four current objectives** of this project are to (1) identify active members of the *Acidobacteria* division that are present in radionuclide and metal contaminated subsurface sediments, (2) compare composition of *Acidobacteria* groups in soils as determined by DNA- or RNA-based methods, (3) attempt to culture novel *Acidobacteria* from soil environments, using a microcapsule approach, (4) obtain whole genome sequence from two *Acidobacteria* division species that belong to phylogenetic groups that are relatively abundant and active in sediment or soil environments.

(1) *Acidobacteria* groups in radionuclide contaminated subsurface environments. Comparisons of the relative abundance and composition of *Acidobacteria* in contaminated subsurface sediments from the NABIR FRC were conducted using PCR amplification of sediment DNA followed by cloning and sequencing. Two subsurface samples from the background area and at least 6 samples from 3 contaminated areas were compared. Both the contaminated and background sediments were dominated by members of the *Proteobacteria, Acidobacteria, Firmicutes, Actinomycetes* divisions, and unclassified species. The *Acidobacteria* comprised 15 to 26% of the total bacterial sequences in the backgrounds sites and 5 to 13% in the contaminated sites.

Within the *Acidobacteria* division, there are currently 8 described subgroups (groups 1 to 8). Through our subsurface sediment study, we have identified at least 7 new subgroups (groups 9 to 15). In the backgrounds sites, the most abundant groups were 1, 3, 4, 5, 6, 10. In the contaminated sites, new groups 9, 10, and 13 were very abundant, in addition to some of the original groups. This result demonstrates a dramatic shift in species composition within the *Acidobacteria* division from the

backgrounds to contaminated sites. Fine scale comparisons of phylotype diversity in the backgrounds and contaminated sites are in progress.

In addition to the FRC surveys, clone/sequence-based surveys of total bacteria and *Acidobacteria* division members were conducted on samples from the Rifle Site, CO and a deep subsurface site at PNNL. At the Rifle site, the *Proteobacteria, Acidobacteria, Firmicutes* and unclassified species were most abundant in clone/sequence libraries. As with most of the surface soils we have characterized, the Rifle site samples primarily contained *Acidobacteria* groups 4 and 6. PNNL site results are in analysis now.

(2) DNA- vs. RNA-based analyses for assessing bacterial community structure. A quantitative, real-time PCR assay for *Acidobacteria* was developed to determine the proportion of *Acidobacteria* relative to total bacteria in environmental samples. We found this assay to work well with surface soils, but had difficulty in subsurface samples from the NABIR FRC due to very low DNA concentrations. We also refined RNA extraction protocols and a reverse transcriptase (RT)-PCR method to assess the presence of 'active' *Acidobacteria* (by interrogation of the RNA) in soil and subsurface sediment samples. A comparative study of total bacteria and *Acidobacteria* division members was conducted in soil using DNA PCR to assess relative abundance of 16S rRNA gene sequences, and using RNA RT-PCR to determine active members (or those with the most ribosomes). Comparisons were made using a T-RFLP method designed for the *Acidobacteria*, and clone/sequence analysis. Sequences from *Acidobacteria* groups 1, 4 and 6 were found to be the most numerous by DNA analysis. However, in the soil tested, groups 1, 3, and 5 were most active by RNA analysis Follow up experiments are focusing on the relative activity and abundance of groups 3, 4, 5, and 6.

(3) Attempts to culture group 6 *Acidobacteria* using Diversa Corporation microcapsule technology. In collaboration with Martin Keller and Karsten Zengler at Diversa Corp., San Diego, CA, and Fred Brockman at PNNL, we are attempting to culture group 6 *Acidobacteria* using Diversa's microcapsule / dilute culture / flow sorting techniques. The first attempt to culture *Acidobacteria* was partially successful, in that we were able to detect them in microcapsules, but were not able to grow them to high titer. Our current attempts involve detection at the microcapsules stage, followed by rolling-circle, whole genome amplification (which can be accomplished from a few to a 100 cells), to generate genomic DNA.

(4) Whole genome sequencing of soil-borne *Acidobacteria*. In collaboration with Peter Janssen, Univ. of Melbourne, Australia, we are providing DNA to the JGI for whole genome sequencing of two *Acidobacteria*, one from group 3 and one from group 4 or 5. These cultures are extremely slow growing and DNA yields are less than 5 μg. The JGI is currently generating libraries from the group 3 isolate.

55

Metagenomic Analysis of Uncultured *Cytophaga* and Other Microbes in Marine and Freshwater Consortia

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

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

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 bulk properties of microbial assemblages, fluorescence in situ hybridization (FISH) assays, 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 the freshwater end of the Delaware Estuary and to the western Arctic Ocean 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 some marine samples, but not for freshwaters nor for a high latitude ocean. High molecular weight DNA from the bacterial size fraction was isolated and cloned into the fosmid vector pCC1FOS (Epicentre). Our libraries consisted of about 5000 clones with an average insert size of 40 kB, representing about 90 genomes, if we assume a genome size of 2 mB.

Screening the libraries revealed several surprises, including genes found previously in metagenomic libraries of oceanic samples. The Delaware River 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. The complete sequence of a fosmid clone containing a *Cytophaga*-like 16S rRNA gene will be presented. 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. Estimates of species-level diversity obtained by rarefaction analysis of fosmid clones bearing 16S rRNA genes differed substantially from clones of PCR products with and without suppression of heteroduplex formation. 16S rRNA gene diversity in the metagenomic library indicated that species-level diversity in this freshwater environment may be on the order of tens of species, much less than current estimates.

We also screened the library for genes indicative of a newly-discovered photoheterotrophic metabolism, aerobic anoxygenic photosynthesis (AAnPS). Marine bacteria carrying out AAnPS contain photosynthesis genes that cluster with those from alpha-, beta-, and gamma- proteobacteria. To date, the diversity and expression of uncultured AAnPS genes in temperate freshwaters have not been examined. We surveyed the Delaware River for *pufL* and *pufM* genes, which encode

AAnPS reaction center proteins, in the fosmid library. Two fosmid clones containing AAnPS photosynthetic operons were completely sequenced and annotated. The operons in the two clones were organized differently than known cultured and uncultured organisms from marine and freshwaters. One clone contained genes most closely related to those of beta-proteobacteria. Preliminary data on *pufM* genes amplified from DNA of Delaware estuary bacteria suggest that most of those genes were most closely related to those of beta-proteobacteria. PCR-amplified and genomic-isolated *pufM* genes cluster separately from currently known cultured and uncultured AAnPS. These data on *pufM* genes in the Delaware estuary indicate an unexpected diversity of estuarine AAnPS bacteria and can be used to explore their ecological success during the transit through the estuary into coastal waters.

56

Approaches for Obtaining Genomic Information from Contaminated Sediments Beneath a Leaking High-Level Radioactive Waste Tank

Fred Brockman[1] (fred.brockman@pnl.gov), S. Li[1], M. Romine[1], J. Shutthanandan[1], K. Zengler[2], G. Toledo[2], M. Walcher[2], M. Keller[2], and Paul Richardson[3]

[1]Pacific Northwest National Laboratory, Richland, WA; [2]Diversa Corporation, San Diego, CA; and [3]DOE Joint Genome Institute, Walnut Creek, CA

The SX Tank Farm at the US Department of Energy's Hanford Site in Washington state was built in 1953 to receive high level radioactive waste, and consists of dozens 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 from beneath the leaking tank were heated in some cases to above 100 degrees Centigrade, contained up to 50 microCuries of Cesium-137 per gram sediment, lower concentrations of other radionuclides, nitrate at 1% to 5% of sediment mass, and pH's to 9.8. These samples are the most radioactive sediments studied to date at the DOE Hanford Site. We hypothesized these extreme conditions would result in a relatively non-diverse community containing novel uncultured microbial divisions.

The original goal was to create a fosmid/BAC library and perform sequencing of clones representing novel uncultured microbial divisions, or less targeted sequencing to characterize the community as a whole. Low biomass levels (10 ^ 5 cells/gram and lower) in combination with very high radioactivity precluded the purification of DNA from the 10's of kilograms of sediment that was required to obtain microgram quantities of DNA for fosmid/BAC library construction. Moreover, extensive characterization of amplified and cloned 16S sequences, from 8 sediments and 30 enrichments from the sediments, failed to show the presence of novel uncultured microbial divisions. Never the less, the 16S rDNA sequencing identified over 40 different genera in the sediments. Approximately 75% of these genera were from the high G+C Gram positive division, highlighting the ability to survive simultaneous extreme conditions for 30-40 years is a widespread trait in this phylogenetic group.

Because the similarity scores of the 16S rDNA clones were mostly >0.95 with known cultured organisms, we elected not to pool enrichments and construct and sequence a fosmid/BAC library. Instead, the project was refocused on (1) an alternative technology for characterizing microbial communities and (2) shotgun sequencing of genomic DNA derived from pooled enrichments.

The first approach involves a novel high throughput microcapsule cultivation method that has been shown to allow culturing of some previously uncultivated microorganisms, allows 16S rDNA characterization, and can be coupled to whole genome (rolling circle) amplification for genome sequencing from the microcolony-containing microcapsules (see poster at this meeting by Zengler et al.). Key aspects of this technology are that it enables propagation of single organisms with extremely slow growth rates and low maximum cell densities, and preserves some of the community interactions and other specific requirements needed for successful cultivation. This approach allows direct access to physiological and genomic information from uncultured and/or difficult-to-culture microorganisms, and is thus fundamentally different than indirect access via shotgun or BAC clones derived from community nucleic acids. Specifically, the units of analysis are living, pure (or nearly pure) microcolonies, as opposed to the disassembled mixture of small fragments of genomes that have lost their biological context in studies using community nucleic acids.

For characterization by the microcapsule culturing approach, the 16 cores were stratified into 4 environmental zones based on levels of radioactivity, temperature, nitrate, and chromium. For each environmental zone and for a nearby uncontaminated borehole, sediments were pooled and cells purified with multiple nycodenz cushion centrifugations. The final cell preparations (from 25 g sediment) contained a total of $3x10^4$ to $9x10^5$ cells by AO counting. The 5 cell preps were each encapsulated to isolate individual cells into microcapsules, the community reconstituted by placing gmd's into a column, and diluted soil extract pumped through the columns to promote slow growth. Three culture conditions were used for each of the 5 cell preps: a high concentration of a mixture of soil extracts under oxic conditions, a high concentration of the same under microaerophilic conditions, and a low concentration of the same under oxic conditions. After several weeks of growth, microcapsules were analyzed by flow cytometry to identify those containing a microcolony, and positives were individually sorted into microtiter wells. A subsample of 2,900 putative microcolony- containing microcapsules from each cell prep were randomly selected for further analysis. The 14,500 cultures were grown further and screened by high throughput FT-IR spectroscopy. Cluster analysis was performed on spectra to identify clusters. Seven to 35 clusters were found per culturing condition per cell preparation from beneath the tank, and 18-90 clusters were found per culturing condition per cell preparation from the uncontaminated sediments. The 16S rRNA gene was sequenced for a representative of each cluster and phylogenetic analysis is ongoing.

The genera obtained by microcapsule culturing will be compared to the approximately 50 isolates previously obtained by plate and liquid culturing. One or more of the most unique microbes (<0.85 similarity to 16S sequences in databases) will be characterized by partial genome sequencing.

For the second approach, the DOE Production Genomics Facility has performed first-pass shotgun sequencing of genomic DNA from enrichments from sediments most highly impacted by the tank waste. Protein hit results will be presented at the meeting.

57

Application of High Throughput Microcapsules Culturing to Develop a Novel Genomics Technology Platform

Karsten Zengler[1], Marion Walcher[1], Imke Haller[1], Carl Abulencia[1], Denise Wyborski[1], Fred Brockman[2], Cheryl Kuske[3], Susan Barns[3], and **Martin Keller**[1] (mkeller@diversa.com)

[1]Diversa Corporation, San Diego, CA; [2]Pacific Northwest National Laboratory, Richland, WA; and [3]Los Alamos National Laboratory, Los Alamos, NM

Project Description

The overall goal of this proposal is to demonstrate the combination of high-throughput cultivation in microcapsules, which gives access to previously uncultivated microorganisms with genome sequencing from one to a few microcolony-containing microcapsules. This will allow direct access to physiological and genomic information from uncultured and/or difficult-to-culture microorganisms. This approach is fundamentally different than characterization and/or assembly of shotgun or BAC clones derived from community DNA or RNA. The units of analysis in our approach are living, pure microbial cultures in microcapsules, as opposed to the disassembled mixture of small fragments of genomes and cellular networks that have lost their biological context in studies using community nucleic acids. It is envisioned that the microcapsule based, high-throughput cultivation method will also be combined with Proteomics technology in the future.

Overall Goal

The overall goal is to prove that microcolonies of previously uncultured microbes derived through this high-throughput cultivation method are sufficient to create genomic sequence information.

The specific goals of this proposed work are:

1. Apply a high-throughput, microcapsule based cultivation technology to capture novel, previously uncultured microbes from a prairie soil relevant to DOE's mission in carbon sequestration.

2. Optimize fluorescent in situ hybridization (FISH) methods to selectively target and sort encapsulated microcolonies of interest using high speed fluorescence activated cell sorting.

3. Employ whole-genome amplification techniques to acquire a sufficient mass of DNA from targeted, encapsulated microcolonies to generate libraries for shotgun sequencing of entire genomes.

4. Develop sensitive methods to amplify specific mRNAs from targeted microcolonies that have been exposed to varying, environmentally relevant conditions.

58

Insights into Community Structure and Metabolism Obtained by Reconstruction of Microbial Genomes from the Environment

Gene W. Tyson[1], Jarrod Chapman[3, 4], Philip Hugenholtz[1], Eric E. Allen[1], Rachna J. Ram[1], Paul Richardson[4], Victor Solovyev[4], Edward Rubin[4], Daniel Rokhsar[3, 4], and **Jillian F. Banfield**[1, 2] (jill@seismo.berkeley.edu)

[1]Department of Environmental Science, Policy and Management, and [2]Department of Earth and Planetary Sciences, and [3]Department of Physics, University of California, Berkeley, CA; and [4]DOE Joint Genome Institute, Walnut Creek, CA

Microbial communities play vital roles in the functioning of all ecosystems. However, the vast majority of microorganisms are uncultivated, thus their roles in natural systems are poorly understood. Random shotgun sequencing of DNA from entire microbial communities is one approach for recovery of the gene complement of uncultivated organisms and for determining the degree of variability within populations at the genome level. Here we report reconstruction of near complete genomes of *Leptospirillum* group II and *Ferroplasma* type II and partial recovery of three other genomes from a natural acidophilic biofilm. This was possible with a modest sequencing effort because the biofilm was dominated by a small number of species populations and the frequency of genomic rearrangements and gene insertions or deletions was relatively low. Because each sequence read came from a different individual, we could determine that single nucleotide polymorphisms are the predominant form of heterogeneity at the strain level. The *Leptospirillum* group II genome had remarkably few nucleotide polymorphisms, despite the existence of a larger pool of low abundance variants. In contrast, we infer that the *Ferroplasma* type II genome is a composite of three ancestral strains that have undergone homologous recombination to form a population of many thousands of unique mosaic genomes. Analysis of the gene complement for each organism revealed the pathways for carbon and nitrogen fixation and energy generation and provided insights into survival strategies in an extreme environment.

59

Growing Unculturable Microorganisms from Soil Communities

Kim Lewis[1] (k.lewis@neu.edu), Slava S. Epstein[1], and Anthony V. Palumbo[2]

[1]Department of Biology, Northeastern University, Boston, MA and [2]Oak Ridge National Laboratory, Oak Ridge, TN

The aim of this project (starting Feb 1, 2004) is to develop methods for growing "unculturable" soil microorganisms. Microorganisms play an important role in shaping the biosphere by affecting biogeochemical cycles and global climate. Novel technologies based on the understanding of microbial life are likely to emerge in the fields of energy production and environmental cleanup, which are of particular interest to DOE. Further development of advanced technologies such as genomics, proteomics, systems biology will be critical to the success of the GTL program. Application of these methods will require access to microorganisms, of which the vast majority, ≥99%, remain "unculturable". Rapid advancement in the understand-

ing of microorganisms and their communities will therefore depend critically on our ability to grow them. The importance of gaining access to uncultivables is underscored by the fact that some of the DOE target sites *have not yielded a single cultivable microorganism*, even though microbial diversity of those environments is substantial. In many locations of the DOE Field Research Center that do produce culturable organisms, these appear not to be the dominant species involved in bioremediation.

The proposed research is based on a method we recently developed that allows for growth of unculturable bacteria by placing them in a simulated natural environment in a diffusion chamber (Kaeberlein, T., Lewis, K., and Epstein, S.S. (2002) Isolating "uncultivable" microorganisms in pure culture in a simulated natural environment. *Science* **296**: 1127-1129). We will use the sites at the DOE Field Research Center as a model to develop methods for growing microorganisms and studying microbial communities. This project will benefit from a considerable amount of information on microbial communities already obtained for these sites.

The **Goals** of this study are:

1. **Characterization of groundwater and soil microbial communities**. An examination of microbial communities will be performed, using molecular approaches. Knowledge of the species composition of these communities will be used to construct specific fluorescent probes. Applying these probes back to the original soil/groundwater samples will allow us to determine the numerically dominant species.

2. **Culturing microorganisms from microbial communities**. Samples of groundwater and soil will be taken from the ORNL sites, and used to culture organisms in a simulated in situ environment in diffusion chambers. Colonies in the chambers will be screened by ARDRA and individual species identified by 16S rRNA approach to verify the presence and growth of species dominating natural communities. We will also isolate organisms from ORNL environments that so far have rendered no cultivable species at all.

3. **In vitro growth of uncultured organisms**. In order to facilitate subsequent analysis, methods to grow uncultured species in vitro will be developed. We find that many isolates will grow on a Petry dish in the presence of another organism from the same environment. We will thus assemble a panel of suitable "helper" organisms that support growth of soil isolates on artificial media. This co-culture approach will serve as a starting point for eventual reconstruction in vitro of a functional soil microbial community, which will be done in our future studies. We have also noted that a majority of uncultured isolates can be adapted to growth in vitro after a number of passages through diffusion chambers. We will use this "domestication" procedure to soil isolates obtained in this study as well.

4. The interesting organisms will be isolated in pure culture and submitted for whole genome sequencing.

60

Gene Expression Profiles of *Rhodopseudomonas palustris* Nitrogenases by Whole Genome Microarray

Y. Oda[1] (yasuhiro-oda@uiowa.edu), S. K. Samanta[1], L. Wu[2], X.-D. Liu[2], T.-F. Yan[2], J. Zhou[2], and C. S. Harwood[1]

[1]University of Iowa, Iowa City, IA and [2]Oak Ridge National Laboratory, Oak Ridge, TN

Rhodopseudomonas palustris is a photosynthetic bacterium that can use many forms of carbon, nitrogen, and electron donors. Under anaerobic conditions it can generate energy from light and convert nitrogen gas to ammonia and hydrogen (a biofuel) by nitrogen fixation. A striking feature of the genome sequence of *R. palustris* CGA009 is genes encoding three different nitrogenases and the accessory proteins needed for nitrogenase assembly. *AnfHDGK*, *nifHDK*, and *vnfHDGK* genes encode iron (Fe)-containing, molybdenum (Mo)-containing, and vanadium (V)-containing nitrogenases. To address the question of how *R. palustris* differentially regulates nitrogenase gene expression, we constructed *anfHnifH*, *anfHvnfH*, and *nifHvnfH* double mutants and analyzed the whole genome gene expression profiles of each mutant and wild-type cells grown under nitrogen-fixing conditions. Each mutant expressed a single functional nitrogenase (Mo, V or Fe) in a minimal medium that contained molybdenum and other trace elements. Wild-type and the Mo-nitrogenase active mutant cells expressed over 150 genes at levels of 2-fold or higher when grown under nitrogen-fixing conditions as compared to when grown with ammonia. Among these were the 30 genes in the *nif* gene cluster, which were expressed at 5- to 200-fold higher (depending on the gene) levels in cells grown under nitrogen-fixing conditions. Genes in the *anf* and *vnf* clusters were not expressed. By contrast, cells with an active Fe-nitrogenase only or an active V-nitrogenase only expressed all of the *nif* genes (except *nifH* which was deleted), all of the *anf* genes, and all of the *vnf* genes. It makes sense that *nif* genes would be expressed because many of them are needed for the assembly of the Fe- and V-nitrogenases. These results indicate that *R. palustris* synthesizes both its Fe- and V-nitrogenases in situations where it is unable to synthesize an active Mo-nitrogenase. The mechanism by which Fe- and V-nitrogenase gene expression is activated is not known, but does not involve relief of Mo repression.

61

Harnessing the Integrative Control of C, N, H, S and Light Energy Metabolism in *Rhodopseudomonas palustris* to Enhance Carbon Sequestration and Biohydrogen Production

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

[1]Ohio State University, Columbus, OH; [2]University of Iowa, Iowa City, IA; [3]Oak Ridge National Laboratory, Oak Ridge, TN; [4]University of British Columbia, Vancouver, BC; and [5]University of California, Los Angeles, CA

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 degradation and metal reduction operate in a single microbial cell. The recently sequenced *Rhodopseudomonas palustris* genome [Larimer et al., Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris*, Nature Biotechnology 22, 55-61, 2004] serves as the raw material for these studies since the metabolic versatility of this organism makes such studies both amenable and highly feasible. A multi-faceted approach has been taken for these studies. On the one hand, novel genes and regulators were identified from investigating control of specific processes by conventional molecular biology/biochemical techniques. In many instances, surprises relative to the role of known regulators, such as the Reg system and CbbR, were noted in *R. palustris*. In addition, a novel phospho-relay system for controlling CO_2 fixation gene expression was identified and biochemically characterized. This latter system, where key regulators contain motifs that potentially respond to diverse metabolic and environmental perturbations, suggests an exquisite means for controlling this key process. Likewise, interesting and important genes and proteins that control sulfur oxidation, nitrogen fixation, hydrogen oxidation, and photochemical energy generation were identified and characterized.

Our studies have shown that the control of CO_2 fixation is integrated and superimposed on the control of nitrogen fixation and hydrogen metabolism in this organism. By interfering with the normal means by which *R. palustris* removes excess reducing equivalents generated from the oxidation of organic carbon, strains were constructed in which much of the electron donor material required for growth was converted to hydrogen gas. The resultant strains were shown to be derepressed for hydrogen evolution such that copious quantities of H_2 gas were produced under conditions where the wild-type would not normally do this. As *R. palustris* and related organisms have long been proposed to be useful for generating large amounts of hydrogen gas in bio-reactor systems, the advent of these newly isolated strains, in which H_2 production is not subject to the normal control mechanisms that diminish the wild-type stain, is quite significant. Moreover, *R. palustris* is unique amongst the nonsulfur purple bacteria in that it is capable of degrading lignin monomers and other waste aromatic acids both anaerobically and aerobically. Inasmuch as the degradation of these compounds may be coupled to the generation of H_2 gas, by combining the properties of the H_2- producing derepressed strains, with waste organic carbon degradation, there is much potential to apply these basic molecular manipulations to practical advances. In addition, our results indicate that *R. palustris* has sophisticated nitrogen acquisition systems that are regulated somewhat differently than in other bacteria. These metabolic processes are driven by light energy, harvested by the photosynthetic apparatus. To maximize this capability, considerable molecular-based study is still

required and the combined expertise of all the investigators of this project, and related projects, is devoted toward this end.

To supplement the more traditional approaches taken above, we have also undertaken a combined microarray/proteomics/metabolomics and bioinformatics approach. Progress has been made towards developing an integrated network of control for the key metabolic processes under study. The whole genome microarrays have given us a global perspective of the changes in metabolic profile that occur under different physiological conditions. Included are metabolic transitions related to the carbon or nitrogen source supplied for growth, the energy source and the gaseous environment. Several genes were shown to be up and down regulated in these experiments, with several implicated in control. These studies have been supplemented by analysis of the proteome under the same growth conditions, using both whole cells and isolated intracytoplasmic membranes [For example, see Fejes et al., Shotgun proteomic analysis of a chromatophore-enriched preparation from the purple phototrophic bacterium *Rhodopseudomonas palustris*, Photosyn. Res. 78, 195–203, 2003]. The end result is that a suite of different, and in some cases, unexpected genes and proteins were identified that respond to specific physiological growth conditions. Moreover, several mutant strains, in which key aspects of metabolism have been altered (see above), were also analyzed by these genomics-based approaches. Beyond merely providing a list of genes and proteins, the transcriptome and proteome screens direct us towards the identification of novel regulators involved in integrating the control of the processes under study.

This multi-faceted approach 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 in the same cell.

62

Gene Expression Profiles of *Nitrosomonas europaea* During Active Growth, Starvation and Iron Limitation

Xueming Wei[1], Tingfen Yan[2], Norman Hommes[1], Crystal McAlvin[2], Luis Sayavedra-Soto[1], Jizhong Zhou[2], and **Daniel Arp**[1] (arpd@bcc.orst.edu)

[1]Oregon State University, Corvallis, OR and [2]Oak Ridge National Laboratory, Oak Ridge, TN

Ammonia-oxidizing Nitrosomonas europaea is a lithoautotrophic bacterium that converts NH^3 to NO_2^- by the successive action of ammonia monooxygenase (AMO) and hydroxylamine oxidoreductase (HAO): $NH_3 + O_2 + 2e^-$ –AMO–> $NH_2OH + H_2O$ –HAO–> $NO_2^- + 5H^+ + 4e^-$. Two of the four electrons return to the AMO reaction and two either provide reductant for biosynthesis or pass to a terminal electron acceptor. The genome of *N. europaea* has been determined and consists of a single circular chromosome of 2,812,094 base pairs. Genes are distributed evenly around the genome, with ~47% transcribed from one strand and ~53% from the complementary strand. A total of 2460 protein-encoding genes emerged from the modeling effort, averaging 1011 bp in length, with intergenic regions averaging 117 bp.

We analyzed the gene expression profile of cells in exponential growth and during starvation using microarrays. During growth, 98% of the genes increased in expression at least two fold compared to starvation conditions. In growing cells, approxi-

mately 30% of the genes were expressed eight fold higher including genes encoding cytochrome c oxidase subunit I, cytochrome c, HAO, fatty acid desaturase and other energy harvesting genes. Approximately 10% were expressed more than 15 fold higher. Approximately 3% (91 genes) were expressed to more than 20 fold their levels in starved cells including the gene encoding multicopper oxidase type 1. Interestingly, the expression of the genes for AMO increased approximately two fold during growth. During starvation, the bulk of the genes were down-regulated with approximately 60% conserving low levels of expression compared to cells in exponential growth. Fewer than 2% of the genes were expressed more than two fold higher in starved cells. Genes expressed during starvation include those encoding NUDIX hydrolase, tyrosinase, multicopper oxidase, lipoxygenase, cycloxygenase-2, a putative transmembrane protein and other oxidative stress genes. Previously we had determined that starved cells transferred to normal medium responded with the induction of global gene expression. We have identified the genes involved in this global response and determined the extent of their expression. We have also identified the genes involved in the adaptation of *N. europaea* to starvation conditions.

Approximately 14% of the coding genes in N. europaea are dedicated to the transport of Fe and to siderophore receptors, yet N. europaea lacks genes for siderophore production (apparently relying on other bacteria to produce them). The growth of *N. europaea* is significantly affected by Fe. When actively growing in normal medium, the cells have a characteristic reddish color (probably due to the accumulation of cytochromes), but in an iron-limited medium, the cells grow poorly and have a lighter color. We carried out a preliminary study to quantify the effect of Fe limitation in the expression of Fe related genes using real-time PCR. Addition of Fe chelators to the Fe-limited growth medium inhibited growth completely for 5 days. *amoA*, *fecR* and *fhuE* were expressed to higher levels in normal growth medium but not in iron-depleted medium. All other Fe-related genes showed no significant expression difference in these treatments. The microarray results showed that Fe-related genes were expressed to higher levels in growing cells than in starving cells. Genes encoding iron transport and binding as well as Fe superoxide dismutase were expressed 10 fold higher in growing cells than in starving cells. The siderophore desferal (produced by *Streptomyces* and other species) promoted the growth of *N. europaea* in iron-limited medium. The gene for the putative desferal receptor (a *foxA* homolog) was expressed to a higher level in iron-limited and desferal-containing cultures than in Fe-containing and desferral-free cultures. The expression of this gene apparently required desferal reinforcing the notion that *N. europaea* grows using the siderophores from other bacteria in Fe-limited environmental conditions. Here we have determined the possible receptor in *N. europaea* to a siderophore from another bacterium.

References

1. Chain, P., J. Lamerdin, F. Larimer, W. Regala, V. Lao, M. Land, L. Hauser, A. Hooper, M. Klotz, J. Norton, L. Sayavedra-Soto, D. Arciero, N. Hommes, M. Whittaker, and D. Arp. 2003. Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph *Nitrosomonas europaea*. *J Bacteriol* **185**:2759-2773.

2. Hooper, A. B., T. Vannelli, D. J. Bergmann, and D. M. Arciero. 1997. Enzymology of the oxidation of ammonia to nitrite by bacteria. *Antonie van Leeuwenhoek* **71**:59-67.

3. Sayavedra-Soto, L. A., N. G. Hommes, S. A. Russell, and D. J. Arp. 1996. Induction of ammonia monooxygenase and hydroxylamine oxidoreductase mRNAs by ammonium in *Nitrosomonas europaea*. *Mol Microbiol* **20**:541-548.

4. Wood, P. M. 1986. Nitrification as a bacterial energy source, p. 39-62. In J. I. Prosser (ed.), Nitrification. Society for General Microbiology, IRL Press, Oxford.

63

Photosynthesis Genes in *Prochlorococcus* Cyanophage

Debbie Lindell[1], **Matthew B. Sullivan***[2], Zackary I. Johnson[1], Andrew C. Tolonen[2], Forest Rohwer[3], and Sallie W. Chisholm[1,4]

*Presenting author

[1]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA; [2]Joint Program in Biological Oceanography, Woods Hole Oceanographic Institution and Massachusetts Institute of Technology, Cambridge, MA; [3]Department of Biology, San Diego State University, San Diego, CA; and [4]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA

Our understanding of the mechanisms of phage-host interactions and their influence on the evolution of both phage and host, is based on a limited set of microorganisms representing an even more limited spectrum of metabolic types. Here we report the presence of genes central to oxygenic photosynthesis in the genomes of three cyanophage from 2 families of double-stranded DNA viruses (*Myoviridae*, *Podoviridae*) that infect the globally abundant marine cyanobacterium, *Prochlorococcus*. The photosystem II (PSII) core reaction center gene, psbA, and one high light inducible (hli) gene type were present in all 3 of the cyanophage genomes. The two myoviruses contain other photosynthesis related genes: One contains the second PSII core reaction center gene, psbD, while the other contains two photosynthetic electron transport genes coding for plastocyanin (petE) and ferredoxin (petF), and both contain additional hli gene types. All of these uninterrupted, full-length genes are conserved in their amino acid sequence with many fewer non-synonymous than synonymous nucleotide substitutions suggesting they encode functional proteins. Phylogenetic analyses indicate that the phage psbA, psbD and hli genes are of cyanobacterial origin, clustering with the corresponding genes from *Prochlorococcus*. They further suggest that these photosynthetic genes were transferred from host to phage multiple times. The phage hli genes cluster with sporadically distributed, multicopy hli types found exclusively in *Prochlorococcus*, suggesting that phage may be mediating the expansion of the hli gene family through the transfer of these genes back to their hosts after a period of evolution in the phage. Such reciprocal evolutionary effects of phage and *Prochlorococcus* on each others photosynthetic gene complement are likely to have significant implications for the success of host and phage in the surface oceans.

64

Metabolomic Functional Analysis of Bacterial Genomes

Pat J. Unkefer[1], Rodolfo A. Martinez[1], **Clifford J. Unkefer**[1] (cju@lanl.gov), and Daniel J. Arp[2]

[1]Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM and [2]Botany and Plant Pathology, Oregon State University, Corvallis, OR

Parallel with the terms genome, transcriptome and proteome, the combined profile of cellular metabolites is the metabolome. Examining changes in the metabolome is a potentially powerful approach to assessing gene function and contribution to phenotype. Achieving the GTL goal of obtaining a complete understanding of cellular function will require an integrated experimental and computational analysis of genome, transcriptome, proteome as well as the metabolome. Moreover, metabolites and their concentrations are a product of cellular regulatory processes, and thus the metabolome provides a clear window into the functioning of the genome and proteome. The profile of metabolites also reflects the response of biological systems to genetic or environmental changes. In addition, metabolites are the effectors that regulate gene expression and enzyme activity. *The focus of this project (Starting February 2004) is the elucidation of gene function by analysis of the metabolome.* We will carry out functional studies using stable isotope labeling and Mass or NMR spectral analysis of low-molecular weight metabolites. Like the proteome, metabolic flux and metabolite concentrations change with the physiological state of the cell. Because metabolite flux and concentration are correlated with the physiological state, they can be used to probe regulatory networks. In prokaryotic organisms, the combination of functional information derived from metabolic flux analysis with gene and protein expression data being developed in other laboratories will provide a powerful approach in identifying gene function and regulatory networks. Our pilot studies will build upon our capability, demonstrate the scientific value, and establish a facility for isotope-enhanced high throughput metabolome analysis of sequenced environmental microbes. Initially, we will study an ammonia-oxidizing chemolithotroph (*Nitrosomonas europaea*). Both play central roles in the global cycles of nitrogen and carbon.

The power of metabolome analysis will be greatly enhanced by applying the combination of stable isotope labeling and mutations. Stable Isotope labeling and NMR/Mass spectral analysis of metabolites will be used to assign metabolic function in three ways. First, we will apply specifically labeled compounds to establish precursor product relationships, and test if putative pathways identified from analysis of the genome are operational. Next, we will develop the capability for functional genomic analysis using comparative metabolomics to reveal the phenotype of a set of so-called silent mutations. This method combines null mutants constructed from the genome sequence by allelic exchange with metabolomic analysis to elucidate the function of unknown ORF's. Finally, we will carry out a full metabolic flux analysis in steady state cultures. Flux analysis will provide input for a stoichiometric model. Many of the advantages of isotope labeling for metabolomics in autotrophs and methylotrophs will be demonstrated throughout this proposal. Once demonstrated, this capability will be even more powerfully applied to heterotrophic organisms growing on complex substrates. These studies will lay the foundation to take similar labeling and metabolomic strategies into the environment to study microbial communities.

65

Genomics of *T. fusca* Plant Cell Wall Degradation

David B. Wilson (dbw3@cornell.edu), Shaolin Chen, and Jeong H. Kim

Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY

This year we have produced a small *Thermobifida fusca* DNA array containing sequences from potential genes involved in plant cell wall degredation and have cloned and characterized a family 10 *T. fusca* xylanase that is induced by growth on xylan but not by growth on cellulose.

The goal of the array study is to identify genes/proteins differentially expressed in *T. fusca* grown on biomass substrates. We will first characterize the transcript profile of *T. fusca* in cellobiose, crystalline cellulose, and xylan cultures using a DNA microarray on a glass slide. Glucose- grown cultures are used as references.

Preparation of DNA Array. Two different *T. fusca* microarray slides are used in this study. One slide contains 123 genes that may be related to biomass degradation as well as 10 "housekeeping" genes as a control (this will be called the123 gene-slide). The other slide has most of the putative *T. fusca* genes identified in its genome sequence (this slide is called genome-slide, hereafter). To select the 123 genes, the *T. fusca* database was searched for putative cellulases and hemicellulases, putative genes that may contain cellulose or chitin binding domains, putative CelR regulated genes, putative secreted proteases/inhibitors, and putative membrane proteins that may be involved in biomass-related signal transduction.

RNA Purification. *T. fusca* was grown in Hagerdahl minimal medium containing glucose, cellobiose, xylan-birchwood, or Solka Floc cellulose. The RNA protect Bacteria regent from Qiagen was used to stabilize RNA in cells after collecting. The phenol/chloroform extraction method (Kieser T. et al. Practical Streptomyces Genetics, pp. 253-363) was initially used for the purification of total RNA from *T. fusca* cells. This method uses buffers containing phenol and thus requires work in a ventilation hood, including the cell-breaking step. It is relatively time-consuming and requires repeated steps of phenol-chloroform extraction and isopropanol precipitation. The RNeasy Midi Kit also was used by following the manufacturer's instruction. This procedure utilizes a centrifugation column to purify total RNA. However, the particulate material in the crude extract can cause contamination and results in impure RNA unable to give labeled cDNA. Therefore, a modified procedure was used, in which phenol/chloroform extraction was applied before column separation.

RNA Labeling. To label the total RNA samples, a protocol from the Institute for Genomic Research (http://www.tigr.org/tdb/microarray/protocolsTIGR.shtml) is used. This procedure labels RNA with aminoallyl labeled nucleotides via first strand cDNA synthesis followed by a coupling of the aminoallyl groups to either Cyanine 3 or 5 (Cy3/Cy5) fluorescent. This indirect aminoallyl labeling procedure yields more uniform labeling and incorporation of the two dyes used in this study.

Hybridization and Scanning. Again the protocol from the Institute for Genomic Research is used for hybridization. Gene Pix4000B is applied to scan arrays. Gene Pix Pro is used for initial data analysis.

Normalization. In order to compare expression levels on the 123 gene-slides, we need to identify a sufficient number of non-differentially expressed genes on each

slide. Therefore, 10 "housekeeping" genes are included. The "Rank-invariant Method" developed by Schadt, EE, et al (In: Feature extraction and normalization algorithm for high-density oligonucleotide gene expression array data. Preprints 303, Department of Statistics, UCLA, Los Angeles, CA) is applied to identify non-differentially expressed genes and to perform normalization. Briefly, the ranks of Cy3 and Cy5 intensities of each gene on the slide are calculated. For a given gene a threshold value d is used for the ranks of Cy3 and Cy5 intensities and a range of I value for the rank of the averaged intensity to determine if a gene is non-differentially expressed. A threshold value of 5 for both d and l will be used for the 123-gene slides. For the genome-slides, the larger number of genes allows us to use a more sophisticated iterative selection scheme as described by Schadt, EE et al. using a program provided by Tseng, GC et al. (*Nucleic Acid Research*, 2001, **29**: 2549-2557).

Other Approaches. In addition, we will use real-time RT PCR and 2-dimensional PAGE to perform further analysis on differentially-expressed genes or proteins.

Xyn10B an endoxylanase from *Thermobifida fusca* was overexpressed in *E. coli* and purified. Mature Xyn10B is a 43 kDa protein that produces xylobiose(SA 95microMol/min/mg) as the major product from birchwood xylan. It hydrolyzes p-nitrophenyl a-D-arabinopyranoside, p-nitrophenyl-b-D-xyloside, and p-nitrophenyl-b-D-cellobioside but at very low rates (<0.1microMol/min/mg). Xyn10B has moderate thermostability, retaining more than half of its xylanase activity after incubation at 55C for 15 hrs and is most active between pH 6-8. Unlike most *T. fusca* hydrolases it has a narrow pH activity profile. Xyn10B is induced by growth of *T. fusca* on xylan or Solka Floc but not on pure cellulose. It does not bind to cellulose, as it lacks a CBM and it appears to be a single domain enzyme.

66

Proteomic Analyses of a Hydrogen Metabolism Mutant of *Methanococcus maripaludis*

M. Hackett[1] (mhackett@u.washington.edu), **J. Amster**[3] (jamster@uga.edu), B. A. Parks[3], J. Wolff[3], Q. Xia[1,2], T. Wang[1,2], Y. Zhang[1], W. B. Whitman[4], W. Kim[4], I. Porat[4], **J. Leigh**[2], and E. Hendrickson[2]

Dept. of [1]Chemical Engineering and [2]Microbiology, University of Washington, Seattle, WA and Dept. of [3]Chemistry and [4]Microbiology, University of Georgia, Athens, GA

Methane-producing archaea catalyze an important step in the anaerobic carbon cycle that converts complex organic matter to CH_4 and CO_2. In total, 1-2 % of all the carbon fixed on earth each year may be processed by the methanogens. In spite of their importance in the anaerobic transformation of complex organic matter, many methanogens are autotrophs and have a very limited capacity to oxidize organic carbon. Thus, the hydrogenotrophic methanogens make the organic components of the cell as well as methane by CO_2 reduction. H_2 is the electron donor for this reaction, but it is too electropositive to couple efficiently with a key step in methanogenesis as well as many of the biosynthetic reactions needed for cellular carbon synthesis. For that reason, methanococci are hypothesized to utilize specialized membrane-bound hydrogenases to generate strong internal reductants from H_2. The *M. maripaludis* genome contains two operons for these energy-coupling, membrane-bound hydrogenases, *eha* and *ehb*. A mutation in *ehb* was constructed by

replacement of a portion of the operon with the *pac* cassette, which encodes puromycin resistance in methanococci. This mutation severely inhibited growth on minimal medium and medium with acetate but not complex medium with amino acids and acetate. This phenotype is a consistent with a role for Ehb in anabolic carbon assimilation.

Proteomic and expression array methods were utilized to further characterize this mutant (S40) compared to the wild type parental strain S2. For proteomic analyses, each strain was grown in two separate cultures, one using ^{14}N and the other ^{15}N nitrogen sources, on medium with acetate. The cells from each culture were then combined into two mixtures: S40 ^{15}N-grown with S2 ^{14}N-grown cells and S40 ^{14}N-grown with S2 ^{15}N-grown cells. Each mixture was then fractionated into soluble and particulate cellular components, resulting in four samples for analysis. Each sample was extracted, proteolytically digested, and run twice on a multidimensional LC-MS-MS system. Peptide identities were determined by computational comparison of collision spectra with the annotated genome sequence, and relative peptide abundances were calculated from the intensities of molecular ion spectra. Protein ratios were calculated based on peptide-to-peptide ^{14}N:^{15}N ratios. Differential protein levels in the mutant vs. the wild type strain were deduced for proteins that had ratios statistically different from 1 in at least two cognate samples, i.e. ^{14}N:^{15}N and ^{15}N:^{14}N for soluble fractions, or ^{14}N:^{15}N and ^{15}N:^{14}N for particulate fractions.

Two enzymes playing central roles in anabolic carbon assimilation were present at lower levels in the mutant compared to the wild type, as supported by differential protein levels for multiple subunits. These enzymes were carbon monoxide dehydrogenase/acetylCoA synthase, and pyruvate oxidoreductase, which catalyze carbon dioxide fixation to acetylCoA and pyruvate, respectively. Each of these anabolic steps is believed to require low potential electrons derived from H_2 via the ehb system. Preliminary expression array data provided addition support for the down-regulation of pyruvate oxidoreductase. These results suggest that, in the absence of low potential electrons provided by Ehb, these anabolic protein levels are down-regulated. However, it is also possible that the levels of these enzymes is affected by the difference in growth rate. In any case, these results eliminates the possibility that these enzyme systems are up-regulated in these mutants.

A variety of proteins were present at higher levels in the mutant compared to the wild type. For example, several ribosomal proteins were more abundant in the mutant, as was the heat shock protein Hsp60. Certain flagellins and flagellum-associated proteins were more abundant in the mutant, and expression array data indicated higher expression levels for flagellin genes. Some subunits of enzymes that catalyze steps in the methanogenic pathway were also present at higher levels in the mutant: these included subunits of methyl-coenzymeM reductase, methyltetrahydromethanopterin-coenzymeM methyltransferase, methylenetetrahydromethanopterin reductase, methenyltetrahydromethanopterin cyclohydrolase, formylmethanofuran-tetrahydromethanopterin formyltransferase, selenium-containing F_{420} reducing hydrogenase, and selenium-containing F_{420} non-reducing hydrogenase. These adjustments may reflect cellular attempts to compensate for the nutritional or growth deficiencies caused by the lack of Ehb activity.

Further work has also been directed to developing new, more rapid proteomic tools to examine the regulation and role of these proteins in carbon assimilation. Currently, we are developing a shotgun method for examining protein expression. In this method, equal amounts of whole cells from cultures grown with 98% ^{15}N are mixed with cells having natural isotope abundance. The cells are lysed in dilute SDS, and the mixture is digested with trypsin. The peptides are fractionated by reverse

phase capillary (150 um ID) HPLC, fractions are collected directly onto MALDI targets, and high-pressure MALDI analyses are performed using a 12 T FTICR mass spectrometer. In the mass spectra, peptides appear as pairs, with one set of peaks from the ^{15}N-labeled cells and one set of peaks from the natural abundance cells. The ratio of the abundance of the two sets of peaks is indicative of the relative expression of the parent proteins. The mass difference between the sets of peaks is equal to the number of N atoms in the peptide. Calculations based upon the genomic sequence indicate that at 5 ppm mass accuracy, 29% of the tryptic peptides (up to 1 missed cleavage) from *M. aripaludis* can be identified solely on the basis of mass. When the numbers of nitrogen atoms are added as a constraint, 48% of the peptides can be uniquely identified. The data collected agrees with these calculations. For a soluble protein extract of wild type cells, the masses of 1184 pairs of peptides were measured. Half of the peptides (503) were uniquely identified to 176 proteins. For comparison purposes, unlabeled proteins obtained under similar growth conditions were examined by 2-D gel electrophoresis (2DGE), and peptide mass fingerprinting was used to identify the 40 most abundant proteins. A majority of the proteins found by 2DGE were among the 176 proteins identified by our new shotgun proteomic technology. Because this method is much more rapid than other proteomic methods and has the potential for high sensitivity and automation, it may substantially reduce the cost of proteomic analyses. At a lower cost, it will be feasible to analyze multiple samples to evaluate the statistical significance of changes in protein expression. This methodology is currently being applied to examine differential display in the membrane protein fractions from S2 and its S40 mutant.

These results reflect significant progress in proteomic and expression array analyses of *Methanococcus maripaludis*, as well as new physiological understanding of the importance of the Ehb hydrogenase. Future analyses will include a comparison of an alternative approach to the processing of the LC-MS-MS proteomic data. Averaging of peptide data for each protein within a sample from a single strain or condition, followed by protein-to-protein comparisons between strains or conditions, may improve proteome coverage while retaining relative quantitative information. The development of the FT-MS methodology in parallel provides the opportunity to validate progress in both methods. In addition, these complementary methods combine the advantages of high proteome coverage by linear ion trap LC-MS-MS and the rapid analytical through put of MALDI FT-MS.

67

Gene Transfer in Hyperthermophiles: *Thermotoga* and *Pyrococcus* as Model Systems

Emmanuel F. Mongodin[1]* (mongodin@tigr.org), Ioana Hance[1], Bruce Weaver[1], Robert T. Deboy[1], Steven R. Gill[1], Tanya Marushak[2], Wei Xianying[2], Patricia Escobar-Paramo[2], Sulagna Gosh[2], Jocelyne DiRuggiero[2], Karl Stetter[3], Robert Huber[3], and **Karen E. Nelson**[1]

*Presenting author

[1]The Institute for Genomic Research (TIGR), Rockville, MD; [2]Dept. of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD; and [3]University of Regensburg, Regensburg, Germany

Whole-genome analysis of the *Thermotoga maritima* genome suggests that 24% of the DNA sequence is most similar to that of archaeal species, primarily to *Pyrococcus* sp. Many of these open reading frames (ORFs) that are archaeal-like are clustered together in large contiguous pieces that stretched from 4 to 21 kb in size, are of atypical composition when compared to the rest of the genome, and share gene order with the archaeal species that they were most similar to. The analysis of the genome suggests that this organism had undergone extensive lateral gene transfer (LGT) with archaeal species. Independent biochemical analyses by Doolittle and workers 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.

In order to investigate the extent of gene transfer across the Thermotogales, we used comparative genomic hybridization (CGH) on ten strains of *Thermotoga* isolated from different locations throughout the world (Table 1). The microarray consisted in 1866 unique PCR products printed in duplicate and representing the whole *T. maritima* MSB8 genome. 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 than10.

Table 1. Strains of *Thermotoga* that were used in the comparative genome hybridization study

Strain	Habitat	Temp (C)
MSB8	Geothermal heated seafloor, Vulcano Island, Italy	55-90
LA4	Shore of Lac Abbe, Djibouti	82
LA10	Shore of Lac Abbe, Djibouti	87
RQ2	Geothermal heated seafloor, Ribeira Quente, the Azores	76-82
RQ7	Geothermal heated seafloor, Ribeira Quente, the Azores	76-82
NE2x_L8B	Naples, Italy	-
NE7/L9B	Naples, Italy	-
S1/L12B	Naples, Italy	-
PB1platt	Oil field at the Prudhoe Bay, Alaska, USA	-
VMA1/L2B	Vulcano Island, Italy	-

Analysis of the CGH data demonstrates that there is a high level of variability in the presence and absence of genes across the different *Thermotoga* strains/species (see Tables 2 and 3). Of the strains that have been compared to the sequenced MSB8, NE2x_LB8, NE_7, 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, 174 ORFs do not have homologs in the MSB8 genome: 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 PB1platt (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. Some of the bigger islands were sequenced in order to determine their size and the gene acquisition/loss in the different *Thermotoga* strains (Table 3). 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.

Table 2. Extent of gene transfer across the *Thermotogales*, based on the CGH results.

***Thermotoga* strains**	RQ2	NE2x_ LB8	NE_7	S1/L1 2B	PB1 platt	VMA1	LA10	RQ7	LA4
% of genes shared with MSB8 [a]	83.5	95.7	88.4	68.6	68.8	12.5	11.0	7.6	6.9
% of genes significantly different from MSB8 [b]	6.9	0.2	6.4	8.0	10.0	69.4	76.6	74.3	83.4

[a] Ratio in CGH was between 1 and 3 ; [b] Ratio in CGH was greater than 10

Table 3. Extent of gene acquisition/loss across the *Thermotogales*, based on the CGH results.

Strain	Region name	Size (kb)	Size in MSB8 (kb)	Deletion/Insertion
RQ2	R1	5	5.75	D
	R2	1	13.16	D
	R5	9.5	9.91	D
	R7	6.5	6.22	??
	R8	4.5	11.9	D
	R9	1.5	12.1	D
	R10	14	9.7	I
	R11	1.5	7.98	D
	R12	9.5	11.98	D
	R13	2	11.53	D
PB1	R1	3	5.75	D
S1	R1	3	5.75	D
NE2X	R1	1.5	1.8	??
	R4	2	2.27	??
	R5	2	2.16	??
	R6	2.5	2.55	??
NE7	R1	??	5.75	??

Thermotoga strains S1/L12B and NE_7, although isolated from the same geographical location, display similar CGH profiles compared to *T. maritima* MSB8. Suppressive subtractive hybridization (SSH) was used to identify sequences that are present in the unsequenced genome of *Thermotoga* strain S1/L12b, but are absent in the strain NE_7. In the first pass of this subtractive study, 61 DNA regions were cloned and sequenced. Using a BlastX analysis, 59 of these clones were matched to genes in strain MSB8, and 3 of these genes (*ligA*, *trpGD*, and an ABC transporter gene) were recognized by more than one clone. To complement the CGH analyses which revealed those genes in strain MSB8 that are missing from several different unsequenced *Thermotoga* strains, future subtractive studies will be used to isolate genes which are unique to these unsequenced strains.

A whole-genome microarray was also constructed for the archaea *Pyrococcus furiosus*. Using custom designed primers, we amplified the 2065 ORFs present in the *P. furiosus* genome. We have a total of 22 new isolates of hyperthermophilic archaea, which we intend to test against the *P. furiosus* array. Seven strains (VB8-I, VB8-II, VB8-V, VB8-VI, VB9-I, VB9-III, and VB11-II) were isolated from Vulcano, Italy by K. Nelson and J. DiRuggiero during a sampling expedition undertaken in 2002. We conducted approximately 25 CGH experiments, and we are in the process of analyzing them. The following archaea strains were isolated from the East Pacific Rise - 13°N-104°W : 12/1, 21/4, 30/2, 30/3, 30/4, 31/2, 32/1, 32/2, 32/3, and 32/4. The following strains were isolated from Juan de Fuca Ridge, East Pacific (North of 13°): JT1, JT3, JT6 and JT10. We conducted approximately 25 hybridization experiments using 5 of the East Pacific strains. A first impression is that the archaeal strains obtained from the Pacific are not as closely related to *P. furiosus* as previously expected, since they hybridize poorly with the *P. furiosus* array. Preliminary results obtained from the hybridizations involving the Vulcano strains will be presented in the poster.

68

Novel Proteins Help Mediate the Ionizing Radiation Resistance of *Deinococcus radiodurans* R1

John R. Battista (jbattis@lsu.edu), Masashi Tanaka, L. Alice Simmons, Edmond Jolivét, and Ashlee M. Earl

Louisiana State University and A & M College, Baton Rouge, LA

Our microarray-based investigations of gene expression in cultures of *D. radiodurans* R1 defined a subset of 33 genes that were induced in response to ionizing radiation (IR) and as cultures recovered from desiccation. Since the process of desiccation and re-hydration introduces DNA damage, we assumed that some of the proteins needed to repair IR-induced damage, including DNA double-strand breaks, would be identical to proteins used to mend DNA damage introduced following desiccation. In other words, the overlap in the cell's response to each stress should specify gene products that directly participate in repair of common DNA damage, potentially identifying novel proteins critical to this process. To test the validity of this assumption, the five hypothetical genes that were induced to highest level in response to each treatment were deleted, and the radioresistance of the resulting strains compared to the R1 parent.

Each of the five genes, which are designated ddrA, ddrB, ddrC, ddrD, and pprA, were replaced with different drug cassettes and the resulting homozygous recessive strains examined for the ability to survive exposure to IR at doses ranging from 1kGy to 13kGy. All mutants were viable and grew with doubling times equal to the parent strain. However, cells of the ΔddrD strain were much smaller than R1, being approximately one third the size of the parent strain. Also, all strains were suitable for natural transformation permitting the uptake and integration of a streptomycin resistance marker into competent cells with efficiencies not different than the parent strain. This result suggests that none of these gene products are required for homologous recombination. The ddrA, ddrB, and pprA mutants exhibited significant increases in sensitivity to IR relative to their R1 parent, indicating that the encoded gene products play fundamental roles in the IR resistance of this species. In contrast the IR resistance of single mutants of ddrC and ddrD do not differ from R1, suggesting that the encoded gene products are either not part of the mechanism that mediates radioresistance, or that they have redundant activities.

In addition to creating the five single mutants, we created double mutants by transforming a Δ*recA* allele into each of the single mutants (ΔddrA, ΔddrB, ΔddrC, ΔddrD, and ΔpprA). We also generated all possible combinations of double mutants using these five alleles in an attempt to establish genetic evidence for potential interactions between the encoded gene products. Analysis of this collection of mutants has revealed: i) that there is a recA-independent pathway that contributes to radioresistance in *D. radiodurans*, ii) that the pprA protein has a novel function that is required for ionizing radiation resistance, but which is not necessary for homologous recombination at least as measured by the cell's ability to carry out allele replacement during natural transformation, iii) that the ddrC and ddrD gene products have complementary activities and that inactivating both proteins is necessary to demonstrate sensitivity to DNA damage, and iv) that the ddrA and ddrB proteins have complementary function and that their activities are most evident when cells suffer high levels of DNA damage.

69

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], George Johnson[1], Derek R. Lovley[2], James K. Fredrickson[3], Wenhong Zhu[4], and John R. Yates III[4]

[1]Argonne National Laboratory, Argonne, IL; [2]University of Massachusetts, Amherst, MA; [3]Pacific Northwest National Laboratory, Richland, WA; and [4]The Scripps Research Institute, La Jolla, CA

Although complete genome sequences can be used to predict the proteins a cell has the potential to express, such predictions do not accurately assess 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.

Proteome analysis generates a variety of data types that must be integrated for efficient assimilation of results. Our project focuses on using electrophoresis methods for protein separation and quantification, coupled with tandem mass spectrometry for protein identification, to determine the patterns of protein expression in a number of microbes pertinent to DOE missions. The data generated by these separation and quantification methods are being integrated by using a suite of World Wide Web applications with a powerful database back-end. The goal is to provide users with a highly interactive web-based resource that contains proteome information, in the context of genome sequence, in formats that enable data interrogations, which will help answer biological questions.

Currently, our protein analyses focus on two microbes with metal-reducing capability: *Shewanella oneidensis* and *Geobacter sulfurreducens*. We have designed and performed experiments in collaboration with GTL projects at the Pacific Northwest National Laboratory (*S. oneidensis*) and the University of Massachusetts (*G. sulfurreducens*). Cells are grown under different conditions to trigger differential protein expression, protein differences are determined by comparative statistical analysis of two-dimensional gel electrophoresis (2DE) patterns, and proteins are identified by tandem mass spectrometry of tryptic digests at the Scripps Institute and Pacific Northwest National Laboratory. This project generates a large volume of data in a variety of formats — including sample descriptions, 2DE images, mass spectra, amino acid sequences, and optical densities — that need to be integrated in a manner that expedites data retrieval and integration. As part of the Microbial Proteome Project, therefore, a suite of four World Wide Web-based databases and interfaces is under development to provide data analysis and integration services at four key access points (described below) in the microbial proteome project work flow.

The Proteomes2 suite of Web applications (http://proteomes2.bio.anl.gov) serves as a Laboratory Information Management System (LIMS) 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 from multiple sites through the Internet. The database currently contains the experimental details for approximately 1400 samples from 11 different microbes

(*Deinococcus radiodurans, Geobacter sulfurreducens, Geobacter metallireducens, Methanococcus jannaschii, Prochlorococcus marinus, Pyrococcus furiosus, Psychrobacter sp.5*, *Rhodopseudomonas palustris, Rhodobacter sphaeroides, Shewanella oneidensis,* and *Synechocystis sp. PCC,*) and links each sample with multiple protein patterns. Over 5000 protein pattern images are currently accessible to authenticated users.

The PMGMS site (http://pmgms.bio.anl.gov/WebSpot/index.htm) allows users to browse through 2DE patterns from which proteins have been digested and analyzed by tandem mass spectrometry for identification based on mass similarity to predicted protein sequences from open reading frame databases. This site also provides views of the mass spectra, the identification results, and the peptide sequences associated with each protein analyzed. Thus, this site integrates the 2DE and peptide mass spectrometry results with gene sequence.

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* and *G. sulfurreducens* experiments are in the process of being incorporated.

GelBank (http://GelBank.anl.gov) currently includes the complete genome sequences of approximately 130 microbes and is designed to allow queries of proteome information. Numerous tools are provided, including the capability to search available sequence databases for specific protein functions and amino acid sequences. Web applications pertinent to 2DE analysis are provided on this site (e.g., titration curves for collections of proteins, 2DE pattern animations). The database is currently populated with protein identifications from the Argonne Microbial Proteomics studies and will accept data input from outside users interested in sharing and comparing results from proteome experiments.

The overall goal of this project is to provide a public resource of protein expression information for microbes in the context of genome sequence.

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

70

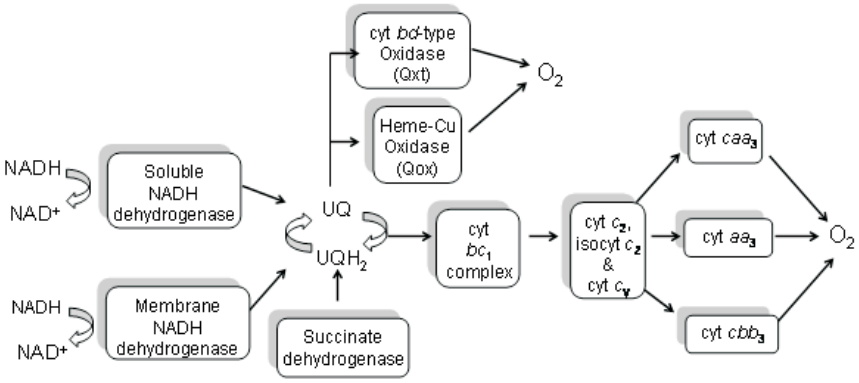
The Molecular Basis for Aerobic Energy Generation by the Facultative Bacterium *Rhodobacter sphaeroides*

Christine Tavano[1], Daniel Smith[2], Matthew Riley[1], Zi Tan[1], Samuel Kaplan[3], Jonathan Hosler[2], and **Timothy Donohue**[1] (tdonohue@bact.wisc.edu)

[1]University of Wisconsin, Madison, WI; [2]University of Mississippi Medical Center, Jackson, MS; and [3]University of Texas Medical School, Houston, TX

The *Rhodobacter sphaeroides* Genomics:GTL consortium seeks to acquire a comprehensive understanding of metabolic pathways, bioenergetic processes, and genetic regulatory networks of a metabolically versatile microbe. This poster reports on ongoing experiments to analyze the different bioenergetic pathways that this facultative bacterium contains to generate energy in the presence of O_2.

The *R. sphaeroides* genome potentially encodes several electron transfer chains that could alter the ability of this facultative bacterium to generate energy in the presence of O_2 (see below).The genome could encode three NADH dehydrogenases, five different terminal oxidases, plus several cytochromes *c* that could transfer electrons between the cytochrome bc_1 complex and three cytochrome *c* oxidases. In addition, two predicted terminal oxidases will utilize quinol as a substrate, thus bypassing the entire cytochrome *c*-dependent branch of the aerobic respiratory pathway.



This part of our project seeks to generate information needed to build testable models of electron flux through branches of these aerobic respiratory pathways at different concentrations of reductant and O_2. In one set of experiments, we are analyzing the ability of individual cytochromes *c* to reduce purified preparations of the major cytochrome *c* oxidases. Initial results indicate that mammalian cytochrome *c* (a mitochondrial counterpart of cytochrome c_2) and *R. sphaeroides* cytochrome c_2 bind the aa_3-type oxidase with similar affinity ($K_m \sim 1.5\text{-}4\ \mu M$), but their V_{max} and ionic strength dependencies are quite different. Mammalian cytochrome *c* has a similar affinity for the aa_3- and the cbb_3 cytochrome *c* oxidases, but the V_{max} values and ionic strength dependencies are different. Comparable experiments with a soluble form of cytochrome c_y and isocytochrome c_2 will allow us to predict which of these proteins are the predominant electron donor to each of these oxidases. In a second line of investigation, we are preparing a series of mutants that each contain only a single

terminal oxidase. In the case of the cytochrome *c*-dependent pathway, strains are also being prepared that demand each oxidase be reduced by either cytochrome c_2 or cytochrome c_y, the two proteins believed to be the predominant electron donors to these enzymes *in situ*. To build testable models for the energy generating capacity of these different respiratory pathways, we will determine the relative bioenergetic capacity of cells containing single routes for electron transfer to O_2.

71

Rhodobacter sphaeroides Gene Expression; Analysis of the Transcriptome and Proteome

Jung Hyeob Roh[1], Jesus Eraso[1], Miguel Dominguez[3], Christine Tavano[3], Carrie Goddard[2], Matthew Monroe[2], Mary Lipton[2], **Samuel Kaplan**[1], and **Timothy Donohue**[3] (tdonohue@bact.wisc.edu)

[1]Department of Microbiology & Medical Genetics, University of Texas Medical School, Houston, TX; [2]Pacific Northwest National Laboratory, Richland, WA; and [3]Bacteriology Department, University of Wisconsin, Madison, WI

The long term goal of the *Rhodobacter sphaeroides* Genomics:GTL Consortium is to engineer microbial cells with enhanced metabolic capabilities. As a first step, we seek to acquire a thorough understanding of energy-generating processes and genetic regulatory networks of this facultative photosynthetic bacterium. In this poster, we will report on first attempts to analyze the transcriptome and proteome of this bacterium when grown under different energy generating conditions. In addition, we will provide a progress report on the analysis of proteins present in purified subcellular fractions or cells grown via aerobic respiration or photosynthesis under low light (3 W/m^2) conditions.

For example, global gene expression patterns of cells grown via aerobic respiration or via photosynthesis at low light (3 W/m^2) conditions indicate that ~60-70% of the ~4600 genes of this bacterium are actively transcribed. Many of these genes show differential patterns of gene expression that are expected based on the changes in energy generation pathways under aerobic respiratory and photosynthetic conditions. In addition, LC/MS-based proteomics of similar cultures has identified >1000 proteins in either whole cells or by analyzing subcellular fractions of known purity. These include soluble and membrane bound proteins and those predicted to be present in one or both of these growth conditions. The poster will summarize the analysis of these genes and the subcellular localization of proteins within aerobic and photosynthetically-grown cells.

72

The Respiratory Enzyme Flavocytochrome c_3 Fumarate Reductase of *Shewanella frigidimarina*

T. P. Straatsma[1] (tps@pnl.gov), E. R. Vorpagel[1], M. Dupuis[1], and D. M. A. Smith[2]

[1]Pacific Northwest National Laboratory, Richland, WA and [2]Whitman College, Walla Walla, WA

S. frigidimarina is a Gram-negative, facultative anaerobe commonly found in marine and freshwater sediments, and is capable to support anaerobic growth using insoluble Fe(III) as terminal electron acceptor. This involves a complex electron-transfer pathway that links primary dehydrogenases in the cell interior with the insoluble, polymeric Fe(III) oxyhydroxides at the surface of the outer membrane. A number of soluble c-type cytochromes are found in the periplasmic space of anaerobically grown *S. frigidimarina*, including a 64-kDa tetra-heme flavocytochrome c_3 fumarate reductase (Fcc_3). The focus of this project is the development and use of computational modeling and simulation tools to characterize complex enzymatic reactivity that includes electron transfer and proton transfer, for which Fcc_3 is taken as our initial target enzyme.

Using density functional theory, we are investigating the relative energies, electronic structure, and optimized geometries for a high- and low- spin ferric and ferrous heme model complex with the hemes in relative conformations as found in classical simulations of the solvated enzyme. The model complex consists of an iron-porphyrin axially ligated by two imidazoles, which model the attachment of the hemes to cytochrome histidines. Using the B3LYP hybrid functional, the doublet ferric heme is found to be lower in energy than the sextet by 8.60 kcal/mol, and the singlet ferrous heme is 7.60 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 5.21 eV, and the low-spin AEA is 5.17 eV. These values are large enough to ensure electron trapping, and electronic structure analysis indicates that the d_π orbital is most likely involved in the electron transfer between neighboring hemes, although the unpaired electron can also occupy a d_{xy} orbital when the imidazole planes are perpendicular. B3LYP geometry optimizations followed by harmonic frequency calculations verified that these conformations (parallel imidazole ligands with a d_π unpaired electron, and perpendicular imidazole ligands with a d_{xy} unpaired electron) are in fact stationary points on their respective *bis*(imidazole) iron porphyrin potential energy surfaces. Calculated imidazole torsion potentials show that, although the torsion potential of the imidazoles of reduced hemes is rather flat, the oxidized hemes have a large barrier to rotation. Calculations of the electron transfer matrix elements for consecutive heme pairs show that the magnitude of the overlap between ET donor and acceptor states, and therefore the electronic coupling, depends strongly on the Fe(3d) hole orbital (d_π vs. d_{xy}), and has implications on the ET pathway among the hemes of Fcc_3.

73

The Cyanobacterium *Synechocystis sp*. PCC 6803: Integration of Structure, Function, and Genome

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

[1]School of Life Sciences, Arizona State University, Tempe, AZ; [2]The Pasarow Mass Spectroscopy Laboratory, University of California, Los Angeles, CA; and [3]The Fellowship for Interpretation of Genomes, Burr Ridge, IL

The cyanobacterium *Synechocystis sp*. PCC 6803 has developed into a model organism for oxygenic phototrophs. Cyanobacteria are very important for the overall carbon balance on earth as they play a major role in global CO_2 fixation, and are thought to be closely related to the endosymbiont in eukaryotes that has given rise to chloroplasts. The Genomics:GTL project on *Synechocystis sp*. PCC 6803 strives to contribute to a comprehensive overview regarding energy metabolism in this cyanobacterium, with structural, metabolomic, genomic, and proteomic contributions. In this abstract, new breakthroughs and insights developed during the past year are summarized, and placed into a perspective of our earlier work.

High-resolution structural imaging. The three-dimensional structure of the *Synechocystis* sp. PCC 6803 cell as analyzed by electron tomography has been refined. Of particular note is that the "thylakoid center", the rod-shaped structure at locations where thylakoids (the internal membrane system carrying the photosynthetic apparatus) converge, can traverse nearly the entire cell. The thylakoid center is likely to have a structural role in keeping thylakoids structurally organized.

Freeze-fracture procedures of cyanobacterial cells recently have been optimized to provide high-resolution scanning electron micrographs in which internal membrane systems are visible and can be followed. Comparison of wild type and specific mutants lacking one or more photosystems, and the recently elucidated three-dimensional crystal structure of the photosystems, may lead to an identification of protein complexes inside thylakoid membranes.

Light microscopic visualization. In vivo protein labeling by means of green-fluorescent protein (GFP) and other fluorescence markers is difficult in *Synechocystis* due to its size and the abundance of highly fluorescent pigments such as phycobilisomes. We have generated fusion constructs with GFP derivatives that enable detection of proteins such as the cell division protein FtsZ.

According to electron microscopic observations, in the absence of significant levels of chlorophyll (as obtained in a mutant where chlorophyll is under light control and cells are grown essentially in darkness) thylakoids appear to be short and disorganized. Interestingly, according to light-microscopic studies, in such systems the localization of fluorescent pigments appears to be very differently organized than in wild type with normal chlorophyll content. Studies regarding the pigment organization and ultrastructure upon chlorophyll synthesis are expected to reveal aspects of thylakoid biogenesis, which thus far has proven to be an enigmatic process.

Carbon metabolism. Targeted deletion mutants have been generated with defects in central carbon metabolism. As suggested earlier, sugar catabolism in *Synechocystis* appears to occur primarily by means of the pentose phosphate cycle, whereas

glycolysis is less active. Methods are now being optimized to detect and quantitate sugar phosphates at reasonably high sensitivity (tens of μM); the next stage will be to follow metabolite fluxes in wild type and mutants of *Synechocystis*.

Synechocystis appears to be very flexible in its metabolism. In terms of its carbon storage compounds, both glycogen and polyhydroxybutyrate can accumulate, depending on the conditions. Thus far, no explanation has been provided for what regulates the nature of the carbon storage compound. Most experimental results we have obtained in this area suggest that the switch regarding the storage compound to be accumulated is under strict redox control.

Cell wall alterations. Cells with impaired cell wall have been generated by targeted gene deletion in order to make the rather small (1-2 μm diameter) *Synechocystis* cells more amenable to light-microscopic investigations to localize pigments and labeled proteins, and to aid in bioenergetic studies on thylakoids. These cells can be maintained under rather isotonic conditions, and their volume can be increased by more than an order of magnitude relative to the wild type.

By impairment of synthesis of the carotenoid myxoxanthophyll by means of targeted deletion mutagenesis, the S-layer (glycocalyx), the outer layer of the cell wall, essentially is removed. Depending on the nature of the gene deletions, membrane transport has been altered, again creating interesting experimental systems for bioenergetics studies.

Proteomics. In order to address concerns over coverage of integral membrane proteins and reproducibility of 2D-gels we are developing a 2D-chromatography system for analysis and quantitation of the membrane protein complexes of *Synechocystis*. By employing a non-denaturing first dimension it is possible to maintain the integrity of the integral membrane protein complexes of photosynthesis and electron transport, as well as their essential cofactors such that we preserve protein/protein interaction information. Intact protein electrospray-ionization mass spectrometry has been successfully integrated into the workflow allowing measurements of integral subunits with as many as eleven transmembrane helices (such as PsaA and PsaB, the 81-83 kDa reaction-center subunits of photosystem I). The current focus is directed at improving the resolution in the first dimension separation while keeping protein complexes intact in order to expand the dynamic range.

High-resolution Fourier-transform mass spectrometry (FT-MS) is being applied to integral membrane proteins for 'top-down' proteomics. Fractions from 2D chromatography are analyzed by electrospray-ionization FT-MS to generate intact protein mass profiles with mass accuracy exceeding 5 ppm and ion isolation with collision activated dissociation to fragment the intact protein directly. Using such techniques it has been possible to sequence through hydrophobic transmembrane domains of proteolipids that remain refractory to other proteomics approaches.

Quantitation remains the critical challenge being faced by proteomics today. We are investigating a modified stable isotope strategy for expression proteomics that will allow measurement of turnover rate and thus proteome flux.

Bioinformatics. The teams at FIG and Argonne National Lab together have constructed a system to support comparative analysis of genomes. The system, called the SEED, now includes the RefSeq data from NCBI. Versions of the SEED extended with newly sequenced versions of genomes from JGI have been prepared and will be made available to any project wishing to use the system. Versions of the SEED running on both Linux and Macintosh systems were demonstrated at SC 2003 in November. During the demonstration a newly sequenced genome was

added to the system, automated annotations were produced, and the ability to share annotations between different versions of the system using peer-to-peer exchanges of data was demonstrated. The system has been used to teach classes in analysis of genomes at both the University of Chicago and Franklin and Marshall College. It is our belief that the system will be used extensively to support annotation efforts, as a framework for exploring genomes in classrooms, and as a central component for construction of integrations of genomic, expression and SNP data. An international collaboration is now forming to establish a framework for rapidly extending the capabilities of the SEED. Three meetings of SEED developers have been planned for the coming year, including two in the US and one in Europe. The first workshop on how to install and use the SEED was held at Argonne National Lab in early January 2004. Versions of the *Rubrobacter* and *Shewenella* genomes were added to the system during the class, automated annotations were generated, and a limited amount of analysis was conducted. We are now responding to suggestions from early users, we will construct a web-based server, we will make DVDs and a straightforward installation procedure available to potential users, and we will offer more classes and workshops to support annotation efforts during 2004.

74

Transport and Its Regulation in Marine Cyanobacteria

Brian Palenik[1] (bpalenik@ucsd.edu), Bianca Brahamsha[1], Jay McCarren[1], Ian Paulsen[2], and Kathy Kang[2]

[1]Scripps Institution of Oceanography, UCSD, La Jolla, CA and [2]The Institute for Genomic Research, Rockville, MD

Cyanobacteria in the open oceans are major contributors to carbon fixation on a global scale. The sequencing and analysis of the genome of marine *Synechococcus* sp. strain WH8102 shows for the first time that these organisms are highly adapted to their oligotrophic marine environment, with relatively small compact genomes and reduced regulatory machinery. The transporters of this organism include ones apparently novel to marine bacteria and ones that are highly conserved across multiple bacterial lineages.

We have shown that a number of putative multidrug efflux transporters can be expressed and function in *E. coli* to increase resistance to multiple antibiotics and related compounds. This provides the first insight into how cyanobacteria may be interacting with other bacteria in natural environments. We have shown that other ABC transporters are required for the unique form of swimming motility in this cyanobacterium and likely function to export the motility apparatus to the outside of the cell. This apparatus has been further defined and shown to include SwmB, the enormous protein encoded by one of the largest known bacterial ORFs. We also have partially characterized organic nitrogen transporters in WH8102 and these are providing insights into the ecology of nitrogen utilization in marine cyanobacteria.

75

Whole Genome Optical Mappings of Two Eukaryotic Phytoplanktons *Thalasiosira pseudonana* and *Emiliana huxleyi*

Shiguo Zhou[1,2], (szhou@lmcg.wisc.edu), Michael Bechner[1,2], Mike Place[1,2], Andrew Kile[1,2], Erika Kvikstad[1,2], Louise Pape[1,2], Rod Runnheim[1,2], Jessica Severin[1,2], Dan Forrest[1,2], Casey Lamers[1,2], Gus Potamousis[1,2], Steve Goldstein[1,2], Mark Hildbrand[3], Ginger Armbrust[4], Betsy Read[5], Diego Martinez[6], Nicholas Putnam[6], Daniel S. Rokhsar[6], Thomas S. Anantharaman[7], and **David C. Schwartz**[1,2,8] (dcschwartz @facstaff.wisc.edu)

[1]Laboratory for Molecular and Computational Genomics, University of Wisconsin, Madison, WI; [2]Department of Chemistry, University of Wisconsin, Madison, WI; [3]Scripps Institution of Oceanography, UCSD, La Jolla, CA; [4]School of Oceanography, University of Washington, Seattle, WA; [5]Biological Sciences, California State University, San Marcos, CA; [6]DOE Joint Genome Institute, Walnut Creek, CA; [7]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI; and [8]Laboratory of Genetics, University of Wisconsin, Madison, WI

Thalassiosira pseudonana and *Emiliana huxleyi* are both eukaryotic unicellular microalgae. *T. pseudonana* is a member of diatoms which is a group of heterokonts found in virtually every water habitat and can be easily recognized by their siliceous cell walls. *E. huxleyi* is a member of chlorophytes, which is found throughout the world's oceans and can be distinguished by its exquisitely sculptured calcium carbonate cell coverings. Diatoms and chlorophytes contribute to the most of the marine primary productivity, and are very important for studying global warming because these organisms together with other phytoplanktons, which make up about 1% biomass on earth, but are responsible for about 50% of the Earth's photosynthesis, therefore, they play a very important role in global carbon circulation between the atmosphere, the ocean, and ocean sediments. Their siliceous or calcium carbonate cell walls can be directly or indirectly used for industry, oil exploration, forensic or biomedical applications. These organisms also play very important roles in the biogeochemical cycling of silica or calcium. There are growing interests for these organisms in the scientific community because these organisms involves the cycling of the basic life elements carbon and oxygen through photosynthesis and respiration and the global climate changes, and also are being used increasingly in a wide range of applications. The optical mapping projects of these two phytoplanktons were carrying out in order to support the DOE genome sequencing projects by providing the whole genome structures and organizations such as chromosome number and ploidy, and providing whole genome restriction map scaffolds for the genome sequence assembly and validations.

The genome *T. pseudonana* was optically mapped using a collection of 65415 single DNA molecules digested by *Nhe* I restriction enzyme. The nuclear genome was estimated to be 34.5 Mb with 24 chromosomes ranging from 339 kb to 3285 kb. This whole genome *Nhe* I map provide us 2752 restriction markers across the genome, which is 1 marker per 12.42 kb DNA on average. With this densely distributed marker map, we were able to align the *in silico* maps from the nascent sequence contigs with our optical restriction marker maps for each chromosome, to find out the orientation and the chromosome assignments for these sequence contigs, and also to determine the gap sizes between the sequence contigs. In return, these processes have greatly speeded up the gap closure and the finishing of the sequence assembly. Furthermore, the homologues of 22 out of the 24 chromosomes can be differentiated at the map level, and some of them have large size variations from sev-

eral tens of kilobases to a few hundreds of kilobases between the homologues of each of these chromosomes. A large inverted duplication, which is about 250 kb, was also detected on one of the chromosome 6 homologues.

The optical mapping of *E. huxleyi* genome has also being carried out from June, 2003. So far, a total of 338,000 singe DNA molecules were digested using *Nhe* I restriction enzyme and collected. Assembly of these single molecule maps resulted in 119 map contigs ranging from 240 kb to 4548 kb. The sum of these map contig sizes is about 221 Mb. As no single map contig looks like finished chromosome map contigs, the accurate estimation of the genome size, structure and organization is still too early. However, one thing is for sure is that the genome of *E. huxleyi* is much larger than originally expected. More collection of the single molecule maps is still needed in order to obtain accurate genome size estimation, and the information about the genome structure and organization.

76

Whole Genome Transcriptional Analysis of Toxic Metal Stresses in *Caulobacter crescentus*

Gary L. Andersen[1] (GLAndersen@lbl.gov), Ping Hu[1], and Harley McAdams[2]

[1]Center for Environmental Biotechnology, Lawrence Berkeley National Laboratory, Berkeley, CA and [2]Department of Developmental Biology, Stanford University School of Medicine, Stanford, CA

Effective bioremediation of metal contaminated DOE sites requires knowledge of genetic pathways for resistance and biotransformation by component organisms within a microbial community. Potentially hazardous levels of lead, mercury, cadmium, selenium and other metals have dispersed into subsurface sediment and groundwater in a number of these sites and represent a challenge for environmental restoration. The aquatic bacterium *Caulobacter crescentus* is an extremely ubiquitous organism with a distinctive ability to survive in low nutrient environments. The association of this unique class of prosethecate bacteria to oligotrophic environments and bacterial biofilms make it an example of an organism that can survive in broad environmental habitats where contamination may be present. We propose to study mechanisms of metal resistance in this bacterium and deduce the regulatory role of selected genes by whole genome transcriptional analysis using high-density microarrays. The *C. crescentus* CB15 genome has been sequenced and is available at: ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Caulobacter_crescentus/. This sequence was used to create a customized 500,000-probe Affymetrix array by the McAdams laboratory at Stanford University that encompasses the entire genome. The expression profile generated by this microarray promises to elucidate metabolic and biosynthetic pathways unique to this organism and to predict conditions in which specific regulatory genes are activated. Our goals in this project are to: 1.) Capture the transcriptome of cells growing in sub-lethal levels of lead nitrate, cadmium sulfate, methylmercury and sodium selenite. 2.) Identify and mutate selected genes involved in survival under increased levels of toxic metals. 3.) Deduce the regulatory role of the *C. crescentus* HU and IHF homologs as well as the *fis* gene using xylose-induced knockouts. Microarray results will be sent to the *Caulobacter* regulatory network database at Stanford University.

Technology Development

Imaging

77

Electron Tomography of Intact Microbes

Kenneth H. Downing (khdowning@lbl.gov)

Lawrence Berkeley National Laboratory, Berkeley, CA

Electron tomography is developing as an effective tool to study subcellular structure at the molecular level. While the thickness of samples that can be studied is limited to a fraction of a micrometer, the resolution is, in principal, sufficient to identify many of the major macromolecular complexes and thus gain insights on their location and interaction. Such information will be essential for the ultimate goals of understanding and building complete computational models of the microbes.

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. The thinner cells are somewhat better suited for study in whole-cell preparations, but we have shown that we can record 2-D projection images by electron microscopy of each of these in frozen-hydrated preparations showing substantial internal detail. We thus retain the native state with no stain or other contrast enhancements, but can see a wealth of internal structures. Frozen-hydrated samples, though, are difficult to work with for several reasons. Aside from the technical issues of specimen preparation and data recording, which are now handled quite routinely, the dense molecular packing and high protein density within bacterial cells makes interpretation difficult for many of the cell components. Large, extended structures, such as cytoskeletal filaments and condensed nucleic acids should be fairly easily discriminated, and once the target resolution range is achieved we expect to be able to identify the major protein complexes.

In the meantime, we have been using a more conventional approach of examining sections of embedded samples. Specimens prepared by high pressure freezing and freeze substitution provide quite good preservation. For example, in eukaryotic samples microtubules provide measure of resolution, and the 40-Angstrom thick protofilaments are often well resolved. This approach is being used to investigate chromium sequestration in *Arthrobacter oxydans* and morphological changes following stress in *Desulfovibrio*, in collaboration with Hoi-Ying Holman at LBNL.

In both frozen-hydrated and embedded samples, we need to develop the ability to identify specific molecular components by labeling. The equivalent of GFP for electron microscopy is a goal of several groups, and several promising approaches are being followed. As a first step, one can use heavy metal cluster labeling of antibodies in the manner as fluorescent antibodies are used at the light microscope level.

Procedures for data collection and processing that overcome the main bottlenecks in generating 3-D representations of the cells have been developing rapidly over the

last few years. Several options now exist for software to control the electron microscope for data collection. To a greater or lesser extent, these relieve the tedium of recording the large number of images required for tomography and the large number of steps needed in collecting each image, thus enabling much faster data collection and ultimately far higher quality data. There is still work that needs to be done to improve the data collection stage, but the remaining rate limiting steps have more to do with visualization of the large data volumes and automated searches through the volume to identify components of interest.

78

Probing Single Microbial Proteins and Multi-Protein Complexes with Bioconjugated Quantum Dots

Gang Bao (gang.bao@bme.gatech.edu)

Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA

To reach the goals of the DOE Genomics:GTL program, there is an urgent need to study individual proteins and multi-protein complexes in microbes. Currently, there is a lack of novel labeling reagents for performing protein intracellular localization and mapping studies. There are few tools that can be used to identify individual proteins and characterize multi-protein complexes in microbial cells, and to visualize and track assembly and disassembly of multi-protein molecular machines. There are no methods to study simultaneous co-localization and dynamics of different intra-cellular processes with high spatial resolution. To meet this challenge, in this study we propose to develop quantum-dot (QD) based strategies for imaging and identification of individual proteins and protein complexes in microbial cells with high specificity and sensitivity. This innovative molecular imaging approach integrates peptide-based cellular delivery, protein targeting/tagging, light microscopy and electron microscopy. Specifically, we propose to develop multifunctional quantum-dot bioconjugates consisting of (1) a quantum dot of 2-6 nm in size encapsulated in a phospholipid micelle, (2) delivery peptides and protein targeting ligands (called adaptors) conjugated to the surface of the QD through a biocompatible polymer. Once the QD bioconjugates are internalized into microbial cells by the peptide, the adaptor molecules on the QD surface bind to specific target proteins or protein complexes that are genetically tagged. Optical imaging will be used to visualize the localization, trafficking and interaction of the proteins, resulting in a dynamic picture but with a limited spatial resolution (~200 nm). The same cells will then be imaged by EM to determine their detailed structures and localize the target proteins to ~4 nm resolution. For each protein or protein complex, selected tags will be tested to optimize the specificity and signal-to-noise ratios of protein detection and localization.

A highly interdisciplinary team has been assembled for this DOE project, with participating faculty members from four universities (Georgia Tech, Emory U, Carnegie Mellon U, and Caltech). The long-term goal is to develop a new multifunctional nanoparticle based molecular imaging platform with enhanced sensitivity, specificity, and spatial resolution. During the proposed three-year period, we will specifically: (1) design, synthesize and characterize quantum dots (QDs) with controlled properties and surface modifications for conjugation with ligands and peptides; (2) conjugate specific ligands such as antibodies and small organic molecules to QDs; (3) develop a peptide-based approach for delivering nanoparticle bioconjugates into

microbial cells; (4) perform fluorescence imaging and electron microscopy to identify, localize, and track proteins and protein complexes. This platform technology will have a wide range of biological and biomedical applications relevant to the Genomes to Life program at DOE, including an improved understanding of multi-protein molecular machines, protein assemblies/networks, and detailed protein functions.

79

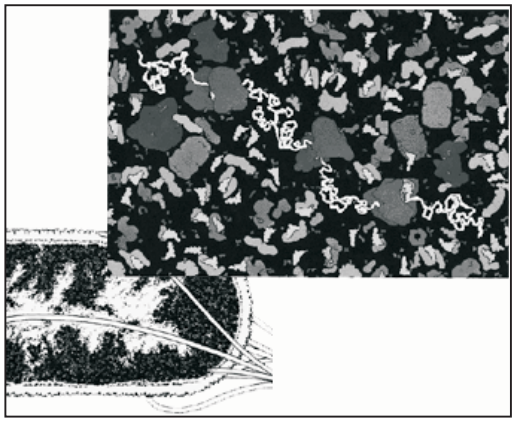
Single Molecule Imaging of Macromolecular Dynamics in a Cell

Jamie H. D. Cate[1-3] (jcate@lbl.gov), Jennifer Blough[1], Hauyee Chang[1], Raj Pai[2], Abbas Rizvi[1], Chung M. Wong[1], Wen Zhou[1], and Haw Yang[1,3] (hawyang@uclink.berkeley.edu)

[1]Department of Chemistry and [2]Department of Molecular and Cell Biology, University of California, Berkeley, CA; and [3]Lawrence Berkeley National Laboratory, Berkeley, CA

We are developing technologies and strategies to image individual proteins and multi-protein complexes in microbes, in order to provide high-resolution and quantitative information on the function of macromolecules in the context of the cell.

The interior of a cell is densely packed with macromolecules. This picture is perhaps best illustrated by a drawing by David Goodsell displayed on the left, showing a number of ribosomal complexes along a RNA chain in the crowded interior of a bacterium[1]. The cellular interior is filled with confined protein molecules and supramolecular complexes that are likely to exhibit different thermal structural fluctuations *in vivo* than those seen *in vitro*[2]. For instance, the diffusion constant of green fluorescent protein (GFP) has been found to decrease four-fold to ten-fold inside a cell relative to diffusion in water[3]. How, then, do the dynamics of biomolecules in crowded environments affect chemical processes in the cell? How do the rates of enzymatic reactions measured *in vivo* compare with those measured *in vitro*?



These questions are very difficult to address by ensemble-averaged assays. A biomolecule inside a cell, constrained in its diffusion, may show a broad location-dependent distribution in its dynamical properties that are distinctly different from those measured *in vitro*. The convoluted spatio-temporal dynamics in cells make it very hard to quantitatively study the various molecular dynamics of a functioning biomolecule. We anticipate that single-molecule spectroscopy, due to its capability of obtaining the individual dynamics from a distribution, will prove invaluable in efforts to unravel how microscopic, molecular interactions impact macroscopic biological functions.

In order to measure macromolecular function and dynamics in the cell, we are developing a single-molecule spectrometer with 3D single-particle tracking capabili-

ties. In an experiment, the biomolecule to be tracked will be conjugated to a tracer element, a surface-passivated nanoparticle that reports the precise location of the biomolecule. In addition to the tracer element, the tracked biomolecule will be labeled with fluorescent probes at strategic sites to allow for simultaneous studies of macromolecular dynamics such as conformational rearrangements, and association and dissociation of macromolecular complexes. We are studying the protein synthesis machinery in *Deinococcus radiodurans* as a model system with which to develop these technologies. Our goal is to establish single-molecule spectroscopy as a general approach to study macromolecules in living organisms.

References

1. Goodsell, D.S. The Machinery of Life (Springer-Verlag, New York, 1998).

2. Ellis, R.J. Macromolecular crowding: obvious but underappreciated. *Trends Biochem Sci* **26**, 597-604 (2001).

3. Elowitz, M.B., Surette, M.G., Wolf, P.E., Stock, J.B. & Leibler, S. Protein mobility in the cytoplasm of Escherichia coli. *Journal of Bacteriology* **181**, 197-203 (1999).

80

Developing a Hybrid Electron Cryo-Tomography Scheme for High Throughput Protein Mapping in Whole Bacteria

Huilin Li (hli@bnl.gov) and James Hainfeld

Biology Department, Brookhaven National Laboratory, Upton, NY

The structures of biological molecular assemblies and their locations inside cells are keys to understanding their functions. Fluorescence microscopy in combination with phase contrast light microscopy is successful in protein localization, but it is limited by its low resolution. This is a serious problem in studying smaller cells such as bacteria with a size of only 1 micrometer. Electron cryo-tomography is an alternative approach to this problem. It provides close to native structure preservation and significantly higher resolution (in the range of 5 to 10nm) three-dimensional structures. However because of the particularly crowded bacterial cellular environment, it is currently difficult to unambiguously identify most proteins. We are developing a hybrid approach, by taking advantage of ultra-structural visualization capability of the cryo-electron microscopy (cryo-EM) and the heavy metal cluster label detection capability of the scanning transmission electron microscopy (STEM) to achieve simultaneously three-dimensional structural visualization and protein mapping. Toward this goal, we will first develop an optimum procedure to label microbial cells, while keeping their structures minimally disturbed. A novel in situ bi-modal tomography protocol of cryo-EM and cryo-STEM will also be developed. To make this method a high throughput tool, universal labels targeted to genetically encoded signatures, such as the Ni-NTA-gold label and 6X-histidine tag system will be developed. The technique will be applied initially to mapping and visualization of the bacterial "cytoskeleton" system and heavy metal resistance protein complexes in *Ralstonia metallidurans*, a microbe of direct DOE interest.

The goals of the project are:

1. To develop a hybrid electron cryo-tomography scheme. We will develop an *in situ* cryo-TEM and cryo-STEM tomography bimodal imaging scheme. The two

tomograms from TEM and STEM tilt series of bacterium embedded in amorphous ice are merged to achieve a simultaneous mapping and visualization of protein complexes in bacteria.

2. To develop a high throughput bacterial labeling strategy based on a 3nm gold particle and genetically encoded signature labeling system, such as Ni-NTA-gold and 6X His-tag proteins. A procedure for mild cell fixing and permeabilization will be developed to allow for label access to the inside of the cell. The procedure is based on the established bacterial cell treatment method in immuno-fluorescence microscopy. After labeling, the bacterial cells will be rapidly frozen in vitreous ice for imaging.

3. To apply the developed methods for high-resolution mapping and visualization of the "cytoskeleton" proteins and multiple heavy metal resistance complexes in *Ralstonia metalliduran*.

81

Probing Gene Expression in Living Bacterial Cells One Molecule at a Time

X. Sunney Xie[1] (xie@chemistry.harvard.edu), Jie Xiao[1], Long Cai[1], and Joseph S. Markson[2]

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA and [2]Department of Chemistry, University of Cambridge, Cambridge, U.K.

We demonstrate for the first time continuous real-time monitoring of gene expression in individual living cells with single-copy sensitivity for a protein. Our approach is based on a modified reporter protein, a short-lived beta-galactosidase (beta-gal), which hydrolyzes a fluorogenic and membrane-permeable substrate and degrades inside a cell within a few minutes. The enzymatic amplification of the hydrolysis product allows us to observe fluorescence bursts corresponding to stochastic expression of the reporter protein in a single *E. coli* cell. Each burst is triggered by the dissociation of the *lac* repressor from the *lac* operator on the *E. coli* chromosome. Moreover, the time traces of the fluorescence bursts exhibit quantized levels corresponding to beta-gal molecules generated and degraded one at a time. The implication of this work to gene expression profiling as well as system-wide studies of gene regulation will be discussed.

Protein Production and Molecular Tags

82

Developing a High Throughput Lox Based Recombinatorial Cloning System

Robert Siegel[1], Nileena Velappan[2], Peter Pavlik[2], Leslie Chasteen[2], **Andrew Bradbury**[2] (amb@lanl.gov)

[1]Pacific Northwest National Laboratory, Richland, WA and [2]Los Alamos National Laboratory, Los Alamos, NM

The selection of affinity reagents (antibodies, single chain Fvs - scFvs) against protein targets can be done using a number of different systems, including phage, phagemid, bacterial or yeast display vectors. Genetic selection methods have also been developed based on yeast two hybrid and enzyme complementation systems. In general, selection vectors are not suitable for subsequent scFv production. Furthermore, once scFvs have been selected, they can be usefully modified by cloning into other destination vectors (e.g. by adding dimerization domains, detection domains, eukaryotic expression in eukaryotic vectors etc.). However, this is relatively time consuming, and requires checking of each individual construct after cloning. An alternative to cloning involves the use of recombination signals to shuttle scFvs from one vector to another. These have the advantage that DNA restriction and purification can be avoided. Such systems have been commercialized in two general systems: Gateway™, uses lambda att based recombination signals, while Echo™ uses a single lox based system to integrate a source plasmid completely into a host plasmid.

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

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

In preliminary experiments we have demonstrated that recombination from one plasmid to another, using 2372 and wild type lox sites and sacB or tetracycline, can occur in vivo at very high efficiency. This opens the possibility of using this system to easily transfer scFvs after selection to other plasmids. However, the utility of this system is not limited to scFvs - any DNA fragment (gene, open reading frame, promoter etc.) can easily be shuttled from one plasmid to another using these lox based signals.

Antibody libraries have been made using these lox sites, and are in the process of being evaluated.

83

Methods for Efficient Production of Proteins and High-Affinity Aptamer Probes

Michael Murphy, Paul Richardson, and **Sharon A. Doyle**

DOE Joint Genome Institute, Walnut Creek, CA

With genome sequencing efforts producing vast amounts of data, attention is now turning towards unraveling the complexities encoded in the genome: the protein products and the cis-regulatory sequences that govern their expression. Understanding the spatial and temporal patterns of protein expression as well as their functional characteristics on a genomic scale will foster a better understanding of biological processes from protein pathways to development at a systems level. Presently, the main bottlenecks in many proteomics initiatives, such as the development of protein microarrays, remain the production of sufficient quantities of purified protein and affinity molecules or probes that specifically recognize them. Methods that facilitate the production of proteins and high affinity probes in a high-throughput manner are vital to the success of these initiatives. We have developed a system for high-throughput subcloning, protein expression and purification that is simple, fast and inexpensive. We utilized ligation-independent cloning with a custom-designed vector and developed an expression screen to test multiple parameters for optimal protein production in *E. coli*. A 96-well format purification protocol was also developed that produced microgram quantities of pure protein. These proteins were used to optimize SELEX (Systematic Evolution of Ligands by Exponential Enrichment) protocols that use a library of DNA oligonucleotides containing a degenerate 40mer sequence to identify a single stranded DNA molecules (aptamers) that bind their target protein specifically and with high affinity (low nanomolar range). Aptamers offer advantages over traditional antibody-based affinity molecules in their ease of production, regeneration, and stability, largely due to the chemical properties of DNA versus proteins. These aptamers were characterized by surface plasmon resonance (SPR) and were shown to be useful in a number of assays, such as western blots, enzyme-linked assays, and affinity purification of native proteins.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, under Contracts No. W-7405-Eng-48, No. DE-AC03-76SF00098, and No. W-7405-ENG-36.

84

Development of Multipurpose Tags and Affinity Reagents for Rapid Isolation and Visualization of Protein Complexes

M. Uljana Mayer, Liang Shi, Yuri A. Gorby, David F. Lowry, David A. Dixon, Joel G. Pounds, and **Thomas C. Squier** (Thomas.Squier@pnl.gov)

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

Our long-term goal is to develop high-throughput methods for the rapid and quantitative characterization of protein complexes in microbial cells *in vivo*. The initial focus of the proposal will be on *S. oneidensis* MR-1, whose metabolism is important in understanding both microbial energy production and environmental remediation. However, these strategies will be applicable to a wide range of microorganisms and will permit the identification of environmental conditions that affect the expression of critical proteins required for the formation of adaptive protein complexes that facilitate bacterial growth. Our hypothesis is that identifying dynamic changes in these adaptive protein complexes will provide important insights into the metabolic regulatory strategies used by these organisms to adapt to environmental changes.

We propose to implement a strategy focusing on the development of multiuse protein tags engineered around a tetracysteine motif (i.e., CCXXCC), which has previously been shown to provide a highly selective binding site for cell permeable arsenic-containing affinity reagents that can be used to first identify and then validate protein complexes in living cells (*Griffin et al., 1998; Adams et al., 2002*). Taking advantage of the large increase in the fluorescence signal associated with binding the proposed fluorescent affinity reagents to the protein tag, it will be possible to use on-line detection to monitor affinity isolation of protein complexes and rapidly identify the proteins in the complex using mass spectrometry. Identification of low-affinity binding interactions in protein complexes is possible by engineering protein crosslinkers onto the bisarsenical affinity reagents. Furthermore, these same protein tags and affinity reagents will permit real-time visualization of steady-state protein abundance and protein-protein interactions, permitting validation of identified protein complexes under cellular conditions and the high-throughput identification of metabolic flow through defined biochemical pathways in response to environmental conditions. Ultimately, these methods will permit an optimization of useful metabolic pathways to fulfill Department of Energy (DOE) goals involving efficient energy utilization, carbon sequestration, and environmental remediation. To accomplish these goals, we propose three specific aims: (1) Identify multipurpose tags with optimized sequences for differential labeling using cell permeable orthogonal fluorescent probes, (2) Optimize expression and *in vivo* labeling of tagged proteins in *S. oneidensis MR-1*, and (3) Develop improved affinity reagents with optional photocrosslinking extensions for stabilizing and identifying cellular protein complexes.

In the next three years we will proceed to fulfill the following three aims:

Aim 1: *Identify Multipurpose Tags with Optimized Sequences for Differential Labeling Using Cell Permeable Orthogonal Fluorescent Tags*. We propose to develop cell permeable reagents that selectively associate with unique tags engineered into bacterial proteins which, in turn, permit highly specific affinity purification strategies for the isolation of protein complexes in *Shewanella oneidensis* and other microbes. These methods are based upon initial results using the fluorescent reagent FlAsH

(Fluorescein Arsenical Helix binder), which specifically interacts with tags containing tetracysteine motifs (i.e., CCXXCC). Optimization of the structure of tags for specific affinity reagents will be accomplished using peptide libraries and computational methods. The structure of the affinity reagents bound to the peptide tags will be determined, allowing for the rational redesign of these affinity reagents to enhance their binding specificities for new affinity reagents.

Aim 2: *Expression and in vivo Labeling of Tagged Proteins in Shewanella oneidensis*. Targeted genes will be cloned under their own promoter into shuttle vectors for *in vivo* expression of tagged proteins following identification of candidate proteins that are expected to form complexes. Expression in *S. oneidensis* MR-1 will be optimized. Tags will be developed to have a minimal impact on cellular metabolism, as determined by measuring their effect on maximal growth rate and molar growth yield in wild-type and modified organisms. Expressed proteins will be identified using fluorescent affinity reagents that recognize specific binding motifs on tagged proteins, permitting the rapid characterization of rates of protein expression and turnover under defined environmental conditions. Optimization of affinity reagents for *in vivo* labeling will involve expression of tagged Aequorea-derived fluorescent proteins (AFP) proteins, whose extent of modification will be measured using fluorescence resonance energy transfer (FRET) methods.

Aim 3: *Develop Improved Affinity Reagents with Optional Photocrosslinking Extensions for in vivo Stabilization and Identification of Protein Complexes*. Purification will involve immobilization of affinity reagents on solid supports. Purified proteins will permit a quantitative characterization of affinity and specificity of affinity labels. Complex purification using both protein encoded tags and bisarsenical probe-based affinity reagents will be optimized. Building upon the scaffolding of known cell-permeable reagents, we propose to develop cleavable crosslinking reagents that stabilize protein complexes and facilitate their isolation and identification using mass spectrometry. Thus, transient protein associations, such as those responsible for fast metabolic control mechanisms, may be identified. Following adduct formation between affinity reagents and protein tags, light activation of the photoreactive moieties will permit crosslinking to binding partners. Following isolation of protein complexes and trypsin digestion, crosslinked peptides will be isolated using an engineered affinity tag, and identified using mass spectrometry.

References

Adams, S.R., R.E. Campbell, L.A. Gross, B.R. Martin, G.K. Walkup, Y. Yao, J. Llopis, and R.Y. Tsien (2002) *New bisarsenic Ligands and tetracysteine motifs for protein labeling in vitro and in vivo: Synthesis and biological applications.* J. Am. Chem. Soc. **124**: 6063-6076.

Griffin, B.A., S.R. Adams, and R.Y. Tsien (1998) *Specific covalent labeling of recombinant protein molecules inside live cells*. Science **281**: 269-272.

85

Development of Genome-Scale Expression Methods

Frank Collart[1] (fcollart@anl.gov), Gerald W. Becker[2], Brian Hollaway[2], Yuri Londer[1], Marianne Schiffer[1], and Fred Stevens[1]

[1]Argonne National Laboratory, Argonne, IL; and [2]Roche Protein Expression Group, Indianapolis, IN

The capability to express proteins in heterologous systems has been an important enabling feature for structural and functional studies of proteins. Although, recent advances in expression technology have significantly increased our capability for the expression of microbial proteins, a significant fraction of proteins encoded by the genome still cannot be expressed in a usable form. We will address these challenging expression problems by application of novel cellular and cell-free technologies to optimize the expression of "insoluble" cytoplasmic and periplasmic proteins. As part of this process, we will evaluate domain-based cloning and expression methods for high molecular weight proteins and putative soluble domains of membrane proteins. The domain-based approach provides an alternative to full length expression and is often used in traditional benchtop approaches. Application of this approach will allow production of soluble domains for many proteins and enable biophysical and biochemical characterization and affinity tag production. These protein resources will support the GTL program and the information gained from these domains may ultimately be used to design a successful strategy for production of the full length protein. Proposed goals include:

1. Extending the capabilities of present high throughput expression platforms to address challenging areas for expression:
   - Apparent insoluble cytoplasmic proteins
   - High molecular weight proteins
   - Soluble domains of membrane proteins
   - Periplasmic proteins
2. Generation of a database using experimental and historical expression data to facilitate development of predictive methods for optimization of expression strategy.
3. Promoting interaction with GTL collaborators to prioritize experimental workflow and facilitate distribution of research resources.

Although automation and high throughput methods can ameliorate some of the cost of protein production, comprehensive protein production strategies will require a balance between optimization of automated methods to enable the cost effective production of clones and proteins and the development of more complex expression strategies for difficult proteins. A major focus of this project is the extension of traditional plate-based methods to address challenging expression problems for proteins from *Shewanella oneidensis* and *Geobacter sulfurreducens*. This dual strategy leverages the cost effectiveness of HT methods to conserve resources and focus on the significant fraction of cellular proteins that remain difficult expression problems but are essential to the undertaking of a system biology approach in understanding microbial cells.

86

Chemical Methods for the Production of Proteins

Stephen Kent (skent@uchicago.edu)
University of Chicago, Chicago, IL

Background

There is a critical need for methods of producing proteins whose existence is predicted by bioinformatic analysis of microbial genome sequence data, in order to undertake their biophysical and functional characterization. Powerful recombinant DNA-based methods exist for the production of proteins in genetically engineered microorganisms or in cell-free translation systems. However, small {<80 amino acid} proteins (~15% of a typical genome) and integral membrane proteins (~25% of a typical genome) have so far proved to be refractory to ready production by these methods. We will prototype novel methods for the high throughput production of milligram amounts of these special classes of proteins using chemical synthesis ['Synthesis of native proteins by chemical ligation.' Dawson, P.E., Kent S.B.H., *Ann. Rev. Biochem*., **69**, 925-962 (2000)].

Our goal is to address the known limitations of *chemical* protein synthesis, based on our intimate understanding of the current state of the art, as exemplified by the total chemical synthesis of the model protein crambin ['Total chemical synthesis of crambin.' Bang, D., Chopra, N., Kent, S.B.H. *J. Am. Chem. Soc.*, In press; 'A one-pot total synthesis of crambin.' Bang, D., Kent, S.B.H., *Angewandte Chemie*, submitted].

Technology Development

The first phase of our research program is focused on the development and optimization of methods aimed at filling the gaps in the tool kit of chemical protein synthesis techniques. These include:

A. Chemical synthesis of peptide-thioesters
   i. Nucleophile-stable thioester-generating resins for solid phase synthesis
   ii. Activation and coupling in the absence of base
   iii. Flow deprotection and cleavage
   iv. High throughput verification of amino acid sequences

B. Ligation at non-cysteine residues
   i. 'Pseudo' native chemical ligation
   ii. Extended native chemical ligation
   iii. 'Traceless' chemical ligation

C. Polymer-supported chemical ligation (solid phase protein synthesis)
   i. Linker chemistries
   ii. Analytical control

In this way, we will develop a practical chemical protein synthesis technology applicable to the rapid preparation of multiple milligram amounts of small and integral membrane protein targets based on predicted gene sequence data. We will then prototype the application of these methods to selected proteins of the model organism *Shewanella oneidensis*.

Significance & Impact

Using the chemical ligation approach the science of chemistry can now be applied, without limitation, to the study of the protein molecule. Chemical synthesis enables the application of all the ingenuity of the modern chemical methods to be applied to the study of the molecular basis of protein function. Applications range from the straightforward replacement of individual amino acid building blocks to much more elaborate and ingenious chemical schemes to engineer new forms of the protein molecule:

- Non-coded amino acids can be incorporated without limitation as to kind, position within the polypeptide chain, and number of substitutions. Non-amino acid building blocks can also be used. For example, a bicyclic β-turn mimetic of fixed geometry was introduced into the HIV-1 protease molecule.[1]
- Post-translational modifications: glycoproteins[2] and glycoprotein mimetics.
- Chemical synthesis can be used to introduce nmr probe nuclei at specific single atom sites in a protein molecule, in any desired number and combination. This can be invaluable for sorting out residue assignments in overlapping regions of the spectra†. Using expressed protein ligation, it is readily possible to mix and match biosynthetically isotope-enriched domains with unlabelled domains in order to simplify the interpretation of nmr spectra of larger proteins.[3]
- Reporter moieties for physical techniques such as EPR or fluorescence spectroscopy can be introduced at will at any desired location within the protein molecule being studied.

Radical re-engineering of the protein molecule has included: building in chemical cleavage sites to unzip the peptide chain at will for protein footprinting[4]; the preparation of proteins containing cyclic polypeptide chains[5]; the construction of topological analogues of proteins (e.g. two N-terminals, no C-terminus; interpenetrating cyclic polypeptide chains[6]).

†This will have particular application to polytopic helical integral membrane proteins; these molecules contain large numbers of identical hydrophobic amino acids in similar chemical environments. Labeling subsets of these residues with nmr probe nuclei will be essential to interpretation of high resolution magic angle spinning nmr spectra of membrane protein preparations.

References

1. Baca M., Alewood P., Kent S.B.H., *Protein Science,* **2**, 1085-1091 (1993).
2. Marcaurelle LA, Mizoue LS, Wilken J, Oldham L, Kent SB, Handel TM, Bertozzi CR, *Chemistry*. **7**, 1129-32 (2001)
3. Cowburn D, Muir TW, *Methods Enzymol.* **339**, 41-54 (2001).
4. Tom W. Muir, Philip E. Dawson, Michael C. Fitzgerald, Stephen B.H. Kent. *Chemistry & Biology,* **3**, 817-825 (1996).
5. Craik DJ, Simonsen S, Daly NL, *Curr Opin Drug Discov Devel.* **5**, 251-60 (2002).
6. Blankenship JW, Dawson PE, *J Mol Biol*., **327**, 537-548 (2003).

87

A Combined Informatics and Experimental Strategy for Improving Protein Expression

John Moult (jmoult@tunc.org), Osnat Herzberg, Frederick Schwarz, and Harold Smith

Center for Advanced Research in Biotechnology, Rockville, MD

The impetus for this project arose out of experience with microbial protein expression in a structural genomics project. We have explored the expression of over 300 non-membrane proteins from *Haemophilus influenzae* and *E. coli* using state of the art over-expression protocols. Our findings are similar to those of other groups: soluble material is obtained for only approximately half of the proteins. In addition to our interest in structural genomics, we are also interested in the *in vitro* and *in vivo* properties of protein molecules. Two questions then naturally arise: what are the relevant differences in properties between successfully expressed proteins and the rest? Further, how can an understanding of these properties by utilized to greatly improve expression success? We will obtain answers to these questions using a combination of informatics and experimental techniques.

A set of approximately 40 proteins already established as spanning all types of expression outcome – plentiful soluble material, insoluble material, no protein expression in healthy cells, and impaired cell growth, form the basis for the experiments. *E. coli* cellular response to over-expression of these proteins will be investigated using full microarray expression profiles. These data will reveal such factors as specific pathways associated with inclusion body formation, up-regulation of proteases and ribonucleases, differential chaperone expression, and previously unsuspected cellular responses. The primary protein properties influencing expression outcome - stability of the folded state and the rate of folding to that state will be investigated, using microcalorimetry and stopped flow measurements. In the third year of the project, we will test hypotheses generated by these experiments, controlling cell conditions as appropriate, and modifying the properties of the test proteins through mutagenesis.

Results from our own and other structural genomics projects will be stored in a publicly accessible database. These data will be mined for factors that affect expression outcome. We have already discovered relationships between protein family size and expression outcome, and between messenger RNA copy number and expression outcome. Other factors to be investigated include the extent of predicted protein disorder, stability, and folding rate. The results of the data analysis will be used to develop tools for predicting likely expression performance and choosing an optimum expression strategy. In addition to making the data publicly available, we will encourage annotation and discussion of the results, and establish a set of 'challenge proteins' – proteins that have so far not been successfully expressed, but which do not fit the emerging model of protein expression outcome.

The outcome of the project will be a set of informatics and experimental strategies. Informatics will provide a synopsis of all relevant information for a protein, ranking alternative strategies for optimization of production. Possible new strategies include the use of GFP and other reporter fusions to monitor up or down regulation of known and newly discovered cell cellular response proteins; utilization of cellular response to control cell growth; protocols for the design of mutants to improve

expression; inhibition of specific proteins shown to affect outcome; and co-expression of proteins found to enhance outcome.

(Funding for this project is about to begin)

88

High-Throughput Production and Analyses of Purified Proteins

F. William Studier[1,2] (studier@bnl.gov), John C. Sutherland[1,2], Lisa M. Miller[3], and Lin Yang[3]

[1]Biology Department, Brookhaven National Laboratory, Upton, NY; [2]East Carolina University, Greenville, NC; and [3]National Synchrotron Light Source, Brookhaven National Laboratory, Upton, NY

Genome sequences allow access to the proteins of an organism through cloning and expression of the coding sequences. Vectors and protocols designed for high-throughput production of proteins in the T7 expression system in *Escherichia coli* are being developed and will be tested by expressing and purifying proteins of *Ralstonia metallidurans*, a bacterium that tolerates high concentrations of heavy metals and has potential for bioremediation. The vectors are designed to accept PCR products and to donate coding sequences for expression as is, with N- or C-terminal tags, or for co-expression with other coding sequences, as with subunits of protein complexes.

Proteins produced from clones are often improperly folded or insoluble. Many such proteins can be solubilized and properly folded, whereas others appear soluble but remain aggregated or improperly folded. Reliable analyses of the state of purified proteins are important for quality assurance in high-throughput production. Stations at the National Synchrotron Light Source analyze proteins by small-angle X-ray scattering (SAXS) to determine size and shape, X-ray absorption microspectrometry to identify bound metals, and Fourier transform infrared (FTIR), UV circular dichroism (CD), linear dichroism (LD) and fluorescence to assess secondary structure and possible intermolecular orientation. An automated sample preparation and loading system to interface between purified proteins in 96-well plates and each of these stations is being constructed to allow high-throughput analyses by these techniques. These assessments of size, shape, secondary structure and metal content of purified proteins will complement analyses such as gel filtration, mass spectrometry and NMR.

Proteomics

89

Ultrasensitive Proteome Analysis of *Deinococcus radiodurans*

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

Department of Chemistry, University of Washington, Seattle, WA

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 have developed the following technologies:

- Fluorescent markers for radiation exposure
- Two-dimensional capillary electrophoresis analysis of the *D. radiodurans* proteome
- Ultrasensitive laser-induced fluorescence detection of proteins separated by capillary electrophoresis

We have generated a number of fully automated two-dimensional capillary electrophoresis separations of proteins extracted from *D. radiodurans*. Figure 1 presents an example, in which the proteins from *D. radiodurans* are first subjected to capillary sieving electrophoretic separation, which is the capillary version of SDS-PAGE using replaceable polymers and which 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. There are 150 components resolved in this separation.

We have developed a fluorescent DNA damage marker for *D. radiodurans*. We have performed the first successful genetic engineering of this organism to express green fluorescent protein (GFP). We have also engineered the organism to express GFP under control of the recA promoter. This gene is expressed in response to DNA damage, and the GFP fluorescence is produced in response to a variety of DNA damage sources. We have demonstrated the production of GFP under this promoter in *D. radiodurans* in response to ultraviolet radiation and toxin (kanamycin and mitomycin C) exposure, Figure 2. We hope to have data on the gamma radiation response of this system by the time of the conference.

Figure 1.

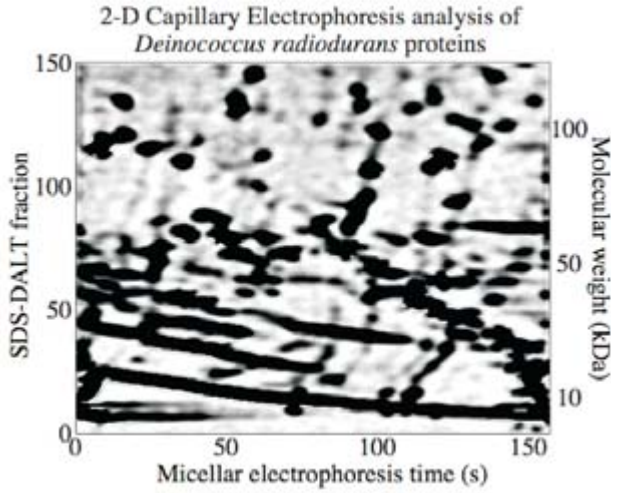
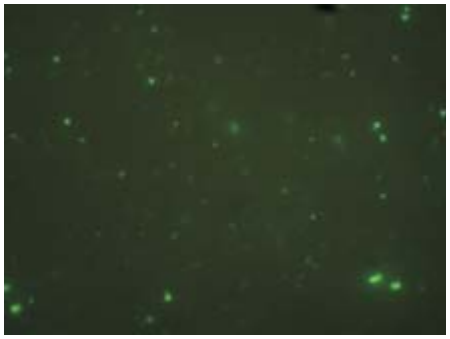


Figure 2. Fluorescent micrographs of recA/GFP engineered D. radiodurans

A. control growth conditions

B. 3-hour exposure to mitomycin C



90

Pilot Proteomics Production Pipeline

Gordon A. Anderson, Mary S. Lipton, Gary R. Kiebel, David A. Clark, Ken J. Auberry, Eric A. Livesay, Vladimir Kery, Brian S. Hooker, Elena S. Mendoza, Ljiljana Paša-Tolić, Matthew Monroe, Margie Romine, Jim Fredrickson, Yuri Gorby, Nikola Tolić, **George S. Michaels** (george.michaels@pnl.gov), and **Richard D. Smith** (dick.smith@pnl.gov)

Biological Sciences Division and Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA

Proteomic analysis of biological samples produces large volumes of data from various mass spectrometric technologies. These datasets allow the identification of peptides and proteins as well as allowing quantization of peptide and protein abundances. This research often requires hundreds or thousands of separate MS experiments. These experiments include a liquid chromatographic separation step coupled with both MS and tandem MS experiments. Data analysis tools are then used to perform database searches in order to identify peptides from tandem MS datasets, interpret and extract detected masses and peak elution times from MS datasets, and assign peptide identifications based on those detected masses and times. These complex multistage analyses require tracking of experimental conditions and sample pedigree. Additionally, quality control analysis needs to be performed at several stages during the process to insure instrument performance and sample preparation quality. Applying MS-based proteomics to the determination of the components present in a sample prepared to specifically contain a given bait protein and its specific binding partners requires additional data tracking and automation. This type of research results in large volumes of data as well as many diverse datasets from carefully designed conditions. Our proteomics production pipeline provides an automation platform to monitor, control, acquire, analyze and organize these results. In order to improve the quality of the research results and record critical experiment and analysis metadata, PNNL has developed an automated proteomics pipeline that includes the following key components:

- **Automated Data Management and Analysis**.

  Sample, analysis, and experimental process data are recorded throughout the pipeline by means of 3 key features. The first is a prototype LIMS system that will cover the experiment design process, gathering process data as well as QA/QC data, and then track those samples until they are ready for MS analysis. Next is commercial freezer monitoring software that will track the movements of specific samples into and out of the freezers, allowing for better inventory control and sample tracking. Barcodes will be used throughout the pipeline to uniquely and quickly identify a sample. Once the samples are ready for MS analysis, they will be handed off to the PRISM system and tracked from there. Analysis of raw mass spectrometer data includes several processing steps involving a combination of commercial data analysis tools and applications developed at PNNL. Automation of the analysis pipeline is performed using our Proteomics Research Information Storage and Management system (PRISM). PRISM automates the capture of data from the mass spectrometers, data reduction of raw data to tables of detected peptides from MS/MS datasets, and tables of detected masses from MS datasets. PRISM then further analyzes this reduced data to develop database tables containing identified peptides and proteins to be used by higher order analysis steps. PRISM allows the users to

monitor the status of analysis and to schedule and track samples through this portion of the proteomics pipeline. These three systems will be connected together through the common use of sample identifiers represented by barcodes.

- **Automation**

  High throughput requires automation; additionally automation provides better control of the process and improves repeatability. Many of the labor intensive and critical aspects of the process are being automated. These automation systems include, protein complex sample preparation, peptide digestion sample preparation, LC-MS and LC-MSMS experiment control. These automation steps include integration with the LIMS to define processing parameters and track the samples as they progress through the system.

- **Data Abstraction Layer (DAL)**

  The DAL is middleware that will provide a level of abstraction for any data storage system in the proteomics pipeline (LIMS, Freezer Software, PRISM, etc). It will provide a generic interface for building tools and applications that require access to the experimental data and analysis results. It will also allow the pipeline data to be extended without making changes in the manner in which an application already looks at the data. For example, it could be used to facilitate a query performed utilizing proteomic data originating from both PNNL and ORNL. The DAL will be used to provide an interface to the pipeline data as required by selected bioinformatics/analysis tools.

Development of the production pipeline lays the foundation for high throughput proteome analysis. This system tracks samples, metadata and raw data for all steps of the process and provides this data to bioinformatics tools through a standard interface. This allows evolution of the PRISM and LIMS system while insulating bioinformatics tools from these changes through the DAL.

91

Characterization of Microbial Systems by High Resolution Proteomic Measurements

Mary S. Lipton (mary.lipton@pnl.gov), Ljiljana Paša-Tolić, Matthew E. Monroe, Kim K. Hixson, Dwayne A. Elias, Margie F. Romine, Yuri A. Gorby, Ruihua Fang, Heather M. Mottaz, Carrie D. Goddard, Nikola Tolić, Gordon A. Anderson, Richard D. Smith, and Jim K. Fredrickson

Biological Sciences Division and Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA

Collaborators: Michael Daly (Uniformed Services University of the Health Sciences), Timothy Donohue (University of Wisconsin), Samuel Kaplan (UT-Houston Medical School), and Derek Lovley (University of Massachusetts)

Developing a systems-level understanding of how cells function requires technologies that are capable of making global measurements of protein abundances (i.e., the "proteome"). At PNNL, new technologies based primarily upon high resolution separations combined with Fourier transform ion cyclotron resonance mass spectrometry have been developed and applied to obtain quantitative and high throughput global proteomic measurements of microorganisms of interest to DOE mission

areas. Among the microorganisms of interest are *Shewanella oneidensis* MR1, *Deinococcus radiodurans* R1, *Rhodobacter sphaeroides* and *Geobacter sulfurreducens*. Significant progress has been made addressing biological questions associated with each of these organisms using high resolution proteomic measurements of cells, and fractions thereof, cultivated under varying conditions.

S. ondeidensis MR-1, a Gram-negative, facultative anaerobe and respiratory generalist, is of interest to DOE because it can oxidize organic matter using metals such as Fe(III) or Mn(III,IV) as the electron acceptors. It can also reduce soluble U(VI) to the insoluble U(IV) form. This ability to reduce U prevents further U mobility in groundwater and subsequent contamination of down-gradient water resources. Microbial reduction shows significant promise for the *in situ* bioremediation of subsurface environments contaminated with U, Tc, and toxic metals such as chromate. A recent revised annotation of the *S. oneidensis* genome suggested a number of changes in the proteins predicted to be expressed by the organism. Using the extensive mass tag database we assembled for this organism and highly stringent criteria for peptide/protein identification, we have for example, analyzed proteome data generated from 172 tryptic digests of *S. oneidensis* MR-1 cellular proteins for the occurrence of peptides associated with proteins less than 101 amino acids in length or that were added to the genome annotation after its initial deposit in Genbank. The mass tag approach also has enabled qualitative experiments to determine the presence or absence of particular proteins in samples as well as quantitative experiments to determine the changes in protein expression upon changes in culture condition. Strategies that use both stable isotope labeling and MS peak intensities of these mass tags provide the basis for quantitation and have been applied to collaborative experiments designed to determine changes in protein expression in cells grown under aerobic and anaerobic conditions.

Similar to *S. oneidensis, Geobacter sulfurreducens* is a dissimilatory metal-reducing bacterium that can reduce soluble U(VI) to insoluble U(IV). Other projects under the DOE Microbial Genome Program have already sequenced the *G. sulfurreducens* genome and have initiated a functional genomics study to elucidate genes of unknown function in this organism. Proteomic efforts with this microorganism are currently focused on creating a mass tag database. Initial global protein expression determinations have shown protein expression in most functional categories as assigned by TIGR. Early uses of the database have centered on determining proteins contained within the membrane of the organism; future studies will be extended to include *Geobacter* dominated microbial communities.

The most significant characteristic of *D. radiodurans* is its ability to resist the lethal effects of DNA damaging agents such as ionizing radiation, UV radiation, hydrogen peroxide and desiccation. The capacity for survival after severe DNA damage at such high levels of ionizing radiation is currently unclear and may be the result of unusually efficient repair and/or protection mechanisms. We utilized the extensive mass tag database developed for *D. radiodurans* and applied a combination of stable isotope labeling and MS peak intensities to determine quantitative changes in protein expression for the organism (1) grown in rich and minimal media, (2) exposed to an acute dose of radiation, and (3) cultured in the presence of chronic radiation.

Rhodobacter sphaeroides 2.4.1 is a-3 purple nonsulfur eubacterium with an extensive metabolic repertoire. Under anaerobic conditions, it is able to grow by photosynthesis, respiration and fermentation. By quantitative measurement of the proteome of *R. sphaerodies* cultured under specific growth conditions, we aim to identify the proteins involved in the different metabolic pathways. For the initial mass tag database, the organism was cultured under both aerobic and photosynthetic conditions,

and differences in the proteins expressed under the two conditions are being determined. Additionally, cellular fractions of these organisms cultured under both aerobic and photosynthetic cell states have been. Photosynthetic cells have been fractionated into 5 relatively discreet fractions (cytosol, periplasm, inner membrane, photosynthetic membrane and outer membrane) and the aerobic cells have been fractionated into 4 relatively discreet fractions (cytosol, periplasm, inner membrane, and outer membrane) in an effort to determine protein localization in the cell. We will be able to determine the changes in localization of specific proteins upon change in cellular state.

The accuracy and precision in which to make proteomic measurements as described above is intricately linked with the instrumentation in which the measurements are made as well as the efficiency of the sample processing methods. Advances in automation of sample processing will reduce variation between digested samples. Additionally, improved methods for quantitation and the application of increasingly sophisticated bioinformatics tools for data analysis will enormously improve the types and quality of the proteomic data available in the future.

92

Advanced Technologies and Their Applications for Comprehensive and Quantitative Microbial Proteomics

Richard D. Smith (dick.smith@pnl.gov), Mary S. Lipton, Ljiljana Paša-Tolić, Gordon A. Anderson, Yufeng Shen, Matthew Monroe, Christophe Masselon, Eric Livesay, Ethan Johnson, Keqi Tang, Harold R. Udseth, and David Camp

Biological Sciences Division and Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA

Essential to realizing the ambitious goals of the Genomics:GTL (GTL) Program is the ability to characterize the broad array of proteins potentially expressed by both individual microbes and complex microbial communities. While recent technological advances are laying the foundation for proteomics approaches that provide more effective, more comprehensive and higher throughput protein measurements, the challenges associated with making truly useful comprehensive proteomics measurements are considerable. Among the challenges are the abilities to identify and quantitate large sets of proteins from highly complex mixtures with components of interest having relative abundances potentially spanning more than six orders of magnitude, that vary broadly in chemical and physical properties, that can have transient and low levels of modifications, and that are subject to endogenous proteolytic processing. Additionally, the utility of proteomics data depends significantly on the quality of the data – both the confidence of protein identifications as well as the quantitative usefulness of the data.

The proteomics technology and approaches developed at PNNL under DOE support employ high resolution nano-scale, ultra-high pressure capillary liquid chromatography (cLC) separations combined with extremely high accuracy mass measurements obtained with Fourier transform ion cyclotron resonance (FTICR) mass spectrometry. The quality of these measurements allow one to identify and designate accurate mass and (cLC) time (AMT) peptide tags that are markers for proteins. These AMT peptide tags can be used in subsequent mass spectrometric measurements, avoiding the throughput limitations associated with routine peptide

identification using tandem mass spectrometry. This approach enables fundamentally greater throughput and sensitivity for proteome measurements. Currently, our prototype FTICR proteomics production line is running in high throughput mode "24/7". A new capability for "data-dependent" tandem mass spectrometry allows otherwise uncharacterized peptides to be selected and characterized directly in the FTICR (i.e., peptides that have not been previously identified and designated as AMT tags). When coupled with stable isotope labeling to allow the direct analysis of two differently labeled samples, this technology identifies those peptides that change significantly in abundance. Additional new developments have 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 to the level approaching that of single cells.

Under DOE support, microbial systems we have extensively characterized include *Deinococcus radiodurans, Shewanella oneidensis, Rhodobacter sphaerodies* and *Geobacter sulfurreducens.* We have developed extensive AMT peptide tag databases for the first two microbes and are in process of developing databases for the latter two. In addition, significant efforts have been made towards characterizing the proteomes of *Rhodopseudomonas palustris*, *Synechocystis*, *Borrelia burgdorferi*, *Desulfovibrio vulgaris*, and *Methanosarcina barkeri*. We have successfully incorporated the use of protein and peptide fractionation in the initial mass tag identification step (based on conventional tandem mass spectrometry in an ion trap), which has increased the dynamic range of these experiments and thus the number of AMT peptide tags. This type of proteomic data can be used in a variety of experiments, ranging from quantitative studies comparing one culture condition to another, to protein localization experiments where cellular fractions are analyzed for their protein content. The use of either stable isotope labeling or MS peak intensities of these AMT peptide tags provides the basis for quantitation. The use of peak intensities potentially circumvents the need for expensive stable isotope labeling methods, and provides a basis for obtaining quantitative information for non-culturable organisms and microbial communities.

A significant challenge for proteomics studies is the immense quantity of data that must be managed and effectively processed and analyzed in order to be useful. Thus, a key component of our program involves development of the informatics tools necessary to make the data more broadly available to the research community and to extract knowledge and new biological insights from complex data sets. A new software tool called VIPER has been developed to automatically process FTICR data sets, which has streamlined data processing. VIPER works with the overall PRISM data management system developed at PNNL to automatically extract and coordinate the use of various types of pertinent information in the application of AMT tags, in addition to managing routine functions, such as FTICR data archiving.

This presentation will describe development and application of new technologies for global proteome measurements that are orders of magnitude more sensitive and faster than previous technologies and that can address many of the needs of the GTL program. The status of the technology will be described in the context of applications, and the basis for extending the applications to more complex microbial communities will also be described.

93

New Developments in Peptide Identification from Tandem Mass Spectrometry Data

William R. Cannon[1] (William.cannon@pnl.gov), Kristin H. Jarman[2], Alejandro Heredia-Langner[2], Douglas J. Baxter[3], Joel Malard[2], Kenneth J. Auberry[4], and Gordon A. Anderson[4]

[1]Statistics and Quantitative Sciences, Computational Sciences and Mathematics Division; [2]Molecular Sciences Computing Facility, Environmental Molecular Sciences Laboratory; [3]Computational Biosciences Group, Biology Division; and [4]Instrument Development Lab, Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA

We present a flexible statistical framework for identification of peptides from the tandem mass spectrometry data. The statistical model is based on a two-sided hypothesis test that compares the likelihood that a spectrum is due to a specific peptide to the likelihood that the peptide arose by chance. The likelihoods are computed from the probability of occurrence of peptide fragments from the parent peptide. These probabilities are empirically derived from fragmentation patterns from a training set of 16,134 spectra of varying charge, composition and length. As a result, a fragmentation model is developed from which model spectra are generated for comparison to real spectra and scoring peptides. The code for the analysis runs on both serial and parallel computers. The statistical model is evaluated on an independent data set of 19,000 spectra using the parallel version of the code on a large Linux cluster.

In addition, we present a sequence optimization approach as an alternative to de novo peptide analysis to reconstruct amino acid sequences of peptides. The sequence optimization can potentially overcome some of the most problematic aspects associated with *de novo* analysis of real MS/MS data such as incomplete or unclearly defined peaks and may prove to be a valuable tool in the proteomics field. We assess the performance of our algorithm under conditions of perfect spectral information, in situations where key spectral features are missing, and using real MS/MS spectral data. The prototype algorithm we use performs well under these situations.

Metabolomics

94

New, Highly Specific Vibrational Probes for Monitoring Metabolic Activity in Microbes and Microbial Communities

Thomas Huser (huser1@llnl.gov), Chad Talley, Allen Christian, Chris Hollars, Ted Laurence, and Steve Lane

Lawrence Livermore National Laboratory, Livermore, CA

We are currently developing a set of new, stable and highly specific intra- and extracellular probes that can monitor metabolic activity inside and in the immediate environment of individual prokaryotic cells. Our sensing technology makes use of vibrational probes (functionalized gold/silver nanoparticles) that monitor the chemical levels inside single microbes with nanometer resolution. These probes consist of specific marker molecules for metabolic byproducts that are chemically linked to the surface of metal nanoparticles with diameters ranging from 40-100 nm. The response of these marker molecules to changes in their local environment can be probed through changes in their characteristic Raman spectrum inside single microbes and in microbial communities. These probes are made of biocompatible and inert materials, they are easy to probe by highly sensitive micro-Raman spectroscopy, and they are very bright and photostable and provide quantitative information about the concentration of metabolic byproducts in their immediate environment. We plan to develop these probes for a range of metabolites and demonstrate their applications in cultured and uncultured microbial communities.

We also demonstrate the use of laser-tweezers Raman spectroscopy, where individual cells are optically suspended in a highly focused laser beam, which at the same time characterizes the chemical activity of the cells by their Raman spectrum. We demonstrate how this capability can be used to distinguish between different cells or monitor their chemical response to external stimuli.

95

New Technologies for Metabolomics

Jay D. Keasling (jdkeasling@lbl.gov), Carolyn Bertozzi, Julie Leary, Michael Marletta, and David Wemmer

Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA

Microorganisms have evolved complex metabolic pathways that enable them to mobilize nutrients from their local environment and detoxify those substances that are detrimental to their survival. Metals and actinides, both of which are toxic to microorganisms and are frequent contaminants at a number of DOE sites, can be immobilized and therefore detoxified by precipitation with cellular metabolites or by reduction using cellular respiration, both of which are highly dependent on cellular metabolism. Improvements in metal/actinide precipitation or reduction require a

thorough understanding of cellular metabolism to identify limitations in metabolic pathways. Since the locations of bottlenecks in metabolism may not be intuitively evident, it is important to have as complete a survey of cellular metabolism as possible. Unlike recent developments in transcript and protein profiling, there are no methods widely available to survey large numbers of cellular metabolites and their turnover rates simultaneously. The system-wide analysis of an organism's metabolite profile, also known as "metabolomics", is therefore an important goal for understanding how organisms respond to environmental stress and evolve to survive in new situations, in determining the fate of metals and actinides in the environment, and in engineering or stimulating microorganisms to immobilize these contaminants.

The goals of this project are to develop methods for profiling metabolites and metabolic fluxes in microorganisms and to develop strategies for perturbing metabolite levels and fluxes in order to study the influence of changes in metabolism on cellular function. We will focus our efforts on two microorganisms of interest to DOE, *Shewanella oneidensis* and *Geobacter metallireducens*, and the effect of various electron acceptors on growth and metabolism. Specifically, we will (1) develop new methods and use established methods to identify as many intracellular metabolites as possible and measure their levels in the presence of various electron acceptors; (2) develop new methods and use established methods to quantify fluxes through key metabolic pathways in the presence of various electron acceptors and in response to changes in electron acceptors; (3) perturb central metabolism by deleting key genes involved in respiration and control of metabolism or by the addition of polyamides to specifically inhibit expression of metabolic genes and then measure the effect on metabolite levels and fluxes using the methods developed above; and (4) integrate the metabolite and metabolic flux data with information from the annotated genome in order to better predict the effects environmental changes on metal and actinide reduction.

Recently, microorganisms have been explored for metal and actinide precipitation by secretion of cellular metabolites that will form strong complexes or by reduction of the metal/actinide. A complete survey of metabolism in organisms responsible for metal and actinide remediation, parallel to efforts currently underway to characterize the transcript and protein profiles in these microorganisms, would allow one to identify rate limiting steps and overcome bottlenecks that limit the rate of precipitation/reduction.

Not only will these methods be useful for bioremediation, they will also be useful for improving the conversion of plentiful renewable resources to fossil fuel replacements, a key DOE mission. For example, the conversion of cellulosic material to ethanol is limited by inefficient use of carbohydrates by the ethanol producer. Identification of limitations in cellulose metabolism and in products other than ethanol that are produced during carbohydrate oxidation could lead to more efficient organisms or routes for ethanol production – metabolomics is the key profile to identify these rate-limiting steps.

Ethical, Legal, & Societal Issues

96

Science Literacy Training for Public Radio Journalists

Bari Scott (bariscot@aol.com)

SoundVision Productions, Berkeley, CA

In the genomic era, journalists bear a greater responsibility than ever to communicate science's rapid advances and their societal implications. The basics of DNA and human genetics, which most journalists still are learning, are no longer enough. Now the media must grasp concepts about the regulation of gene expression, the activity of proteins, the workings of RNA and other mechanics of the cell, both in humans and other forms of life such as microbes. Advances in these areas have profound implications, and journalists are obliged to provide timely and accurate information to the public.

SoundVision Productions®, creator and facilitator of two successful week-long workshops for public radio journalists in 1999/2000 and 2001/2002, will develop a new series of three, week-long science literacy training workshops and related activities for public radio reporters and producers. At each of the three workshops, twelve mid-career public radio producers and reporters, selected through a competitive process, will be introduced to cutting-edge research; encouraged and given the tools to report on science stories and their corresponding ethical considerations; and trained in the protocols of science journalism. Based on past experience, SoundVision anticipates that participants will represent stations and national/regional programs that reach broad audiences. The recruitment and selection process will also ensure ethnic, racial, and gender diversity, with particular emphasis on including journalists from rural and minority-controlled stations and networks.

We will explore the interactions among DNA, RNA, proteins, and the overall complex machinery of the cell, and teach reporters what scientists are learning about the most basic elements of life. Applications in the areas of environment, energy, and health will be addressed. A key workshop focus will be post-human-genome-project ethics, such as new questions in the science-business relationship, the impact of highly patented science, and the risks and responsibilities of attempting to manipulate life.

Each workshop will center around 15 to 20 presentations by scientists, science journalists, science researchers, and radio production professionals. Sessions will orient producers to basic science, focus on the craft and responsibility of science journalism, and explore techniques for presenting complex scientific content on radio. This third component is particularly important due to the specific production needs that distinguish radio from other media. Each workshop also will include a field trip and several informal gatherings with scientists to develop relationships and learn more about their ideas and research.

The week-long workshops will be held in key U.S. locations in order to reach more producers throughout the country. The first will be held in Boston, co-hosted by WGBH-FM and The Whitehead Institute. The second workshop will be held in San Francisco at KQED-FM, and the third will be in Austin, Texas in cooperation with Latino USA/KUT-FM at the University of Texas.

SoundVision's training methodology is one that has garnered positive results in previous cohorts of public radio journalists. Even years after attending, participants from small rural to large metropolitan stations report that the week-long workshops in 1999 and 2001, which were funded by the Department of Energy, still have significance to their work. Producers and reporters continue to benefit from their familiarity with the basics of DNA research, an ability to identify stories that they wouldn't previously have tackled, and skills in getting behind press releases and scientific papers in order to create compelling public radio features. These innovative workshops provide participants with methods that improve their confidence and their ability to communicate complex and emerging scientific research to the audience. The goals of these workshops are: to increase the number and quality of science stories produced for radio; to add to the number of reporters able to competently report on complex research processes, discoveries, and implications; and ultimately, to help minimize the currently widening knowledge gap between the scientific community and the public. We believe that their collective work on all geographical levels will help lead to a better public understanding of current scientific research and its social implications.

The project also includes a website that highlights transcripts and selected audio from the training sessions, "tip sheets," and online resources available to participants and interested users. There will be follow-up teleconferences to support participants in pursuing complex and rewarding science stories for their communities. In addition, if funds permit, a pilot interactive DVD with highlights of the workshop will be offered to rural and minority-controlled stations and networks. If successful, the "pilot" may be developed and used for other applications.

As in our previous workshop projects, a comprehensive evaluation will be conducted by Rockman ET AL, a well-established, San Francisco-based evaluation firm with expertise in evaluating media projects and in assessing the impact of training on journalistic practice.

Although the project targets public radio producers and reporters, with slight modifications the workshop is applicable and adaptable for training news directors and editors; live interview, call-in hosts and their producers (since so much air time is dedicated to that format); and even television staff.

97

The DNA Files

Bari Scott (bariscot@aol.com)

SoundVision Productions, Berkeley, CA

The Human Genome Project has yielded more than simple sequence data. Scientists now have better lab technology, more sophisticated informatics, and mountains of new data to explore. Each day they discover more about the remarkably intricate workings of cells and the interdependence of living organisms. Researchers have also begun to chart genetic variation among humans, study the interconnected systems of biology, explore and manipulate the most basic elements of life, and exploit biological processes in industry and healthcare. All of these advances have benefited from a genome-based analytical framework, which incorporates data and insights from many genome projects, including microbial data as well as model organism and human genome sequence information.

This new frontier in science is immensely promising and extremely challenging – for both scientists and society. It requires us to assess risk, weigh risk against benefits, reconsider our definitions of health, normalcy, and appropriate treatment, and view human responsibility in the context of an active, complex physical environment. It raises questions about intellectual property, marketing and commercialization, and the legal and ethical roles of government and corporations. It may even require us to rethink our relationship to our environment and our understanding of the basic components of life.

We believe it is essential for society to learn more about the concepts and approaches underpinning genomics and cellular systems in order to prepare for the discoveries, medicines, and technologies likely to emerge from this new way of thinking about science. Citizens currently have little idea what questions scientists are asking, what innovations might result, and what ethical and legal challenges scientists face.

SoundVision Productions®, creator of the highly acclaimed, nationally distributed public radio documentary series *The DNA Files*, will produce, market, distribute, outreach, and evaluate one hour-long documentary, one five-minute feature, web articles and on-line resources about the scientific, ethical, legal and societal issues raised by the *Genomics:GTL* initiative. These make up one topic-related ensemble for *The DNA Files 3*, a new series of five hour-long radio documentaries, accompanying five-minute features, a multimedia web site, and promotional materials that will inform a diverse public about the complex changes being brought about by advances in genomics and systems biology. “Ethics Beyond the Genome” will alert the public to some of the most important ideas and challenges emerging from systems biology and offer the public intellectual tools to participate in legal and social policy debates about our technological future. The four other topics in the series are: “Our Common Genes: Bugs, Mice and the Human Body”; “Toxicogenomics and Individual Variation”; “The RNA World and Immunology”; “Neurobiology and Our Genes”.

Each of the new topics represents a rapidly developing field within genomics rarely covered in depth by the media. The project will help disseminate the science of genomics by highlighting recent findings and integrating these with examples of ongoing research. *The DNA Files* content will add to public awareness, knowledge

and understanding by providing a foundation in science and the scientific method, an introduction to those engaged in scientific pursuit, and a sampling of ethical, legal and societal issues. These elements will offer audiences an awareness of the societal benefits of research and the intellectual tools to join in legal and social policy debates.

The earlier two DNA Files series won many honors, including the coveted Peabody, DuPont Columbia, and AAAS awards. SoundVision promotes excellence through in-depth reporting, content development with diverse audiences in mind, targeted distribution and marketing, and ongoing evaluation. Its systematic journalistic methodology ensures accuracy and appropriate context through a reliance on traditional reporting techniques; extensive background research, producer training, reporting plans, and script reviews; plus regular consultations with scientific advisors. We also design materials to have lasting relevance by focusing on underlying trends rather than the latest press release.

SoundVision will employ the experience of its project team, and evaluation, to develop content that will engage the audience and incorporate their interests. SoundVision's experienced producers employ creative sound (such as audio montage and archival footage) and content, explaining basic research and related concepts through dramatization, description and other methods that both engage and inform listeners and web users. The materials from our programs, in-depth articles and an extensive resource library add to the infrastructure for science education through their long-term availability to educators and the public on www.dnafiles.org. As finances permit, staff updates shows, offers them for rebroadcast, and upgrades the Web site.

The DNA Files I and *II* were each featured on 200 public radio stations, including in major U.S. broadcast markets. The 2005 series will expand this strong base further by working with community radio and production entities that reach ethnic minority and rural audiences. NPR is interested in distributing the series again.

We will refine and evaluate the impact of our content through focus groups and direct observation of user activities on the web. We continue to update portions of our shows and redesign, maintain and add to www.dnafiles.org. Previous evaluations have demonstrated to us that users view the site as trustworthy and continue even to rely on material updated from the original 1998 documentaries. A nonprofit, 501(c)(3) organization, SoundVision has been producing and disseminating informational programming relating to the humanities, science and technology since 1995, focusing on the presentation of complex subject matter to a general audience.

Much of the key staff from the earlier DNA Files, including host John Hockenberry, have committed to a new series. The Department of Energy provided initial funding for the original *The DNA Files*, which has since been supported by a variety of foundations and agencies.

Unindexed Late Abstracts

The Environmental Molecular Sciences Laboratory:
Application to Biology and Biological Grand Challenges

J. W. Rogers
EMSL Director, Pacific Northwest National Laboratory, (509) 376-5328, jw.rogers@pnl.gov,
Richland, WA

The William R. Wiley Environmental Molecular Sciences Laboratory (EMSL) is a national scientific user facility located at Pacific Northwest National Laboratory (PNNL) in Richland, Washington. The EMSL was developed and is operated for the DOE as a multiprogram national user facility for molecular studies focused on solving the major environmental and biological problems facing DOE and the Nation. The EMSL http://www.emsl.pnl.gov/ has signature characteristics which include: integration of theory, modeling, and simulation, with experiment; multidisciplinary teams and collaborative modes of operation to solve major scientific problems; teams responsible for development of extraordinary tools and methodologies; scientists who design experimental strategies and operate state-of-the-art instruments; integrated operation and coordinated execution; education and training in the use of sophisticated instrumentation / computation / systems and approaches; a cyber infrastructure that facilitates productive remote interactions; and the charter to deliver capability in a transparent manner and to facilitate user outreach. The EMSL also boasts unparalleled resources and infrastructure in high-performance computing and informatics (e.g., Linux cluster supercomputer which supports computational biology and bioinformatics), nuclear magnetic resonance spectroscopy (e.g., 12 NMR spectrometers (300-900 MHz) and one pulsed EPR spectrometer), multimodal optical spectroscopies and imaging technologies (lasers; magnetic resonance; atomic force, near-field optical, and other scanning microscopies; electron microscopies), as well as advanced mass spectrometers for global proteomics (four Fourier transform ion cyclotron resonance mass spectrometers, Sciex QSTARR quadrupole time-of-flight mass spectrometer, Five Finnigan LCQ ion trap spectrometers, and a Finnigan TSQ 7000 triple quadrupole spectrometer), all of which are applicable to biological research. Supporting these technologies are world-class staff at PNNL, including biologists, chemists, physicists and software engineers, as well as instrument designers and builders.

The EMSL is now seeking concept papers for a scientific grand challenge in the area of membrane biology. Successful proposals will pose scientific questions that cannot be readily addressed without access to the full range of scientific instrumentation, computational resources, and research teams located at the EMSL for substantial periods of time. They are expected to attract and involve some of the best research scientists in the area chosen for study. Research areas of interest include biological membrane processes in cells (e.g., energy transduction, photosynthesis, signal transduction, dynamics of membrane proteins, and regulation of conformation states of proteins). Understanding membrane processes requires a systems-level analysis of fundamental cellular processes, the characterization of which is particularly well suited to the capabilities of the EMSL. The EMSL requests submission of a concept paper outlining the grand challenge goal, approach, technical requirements, and expected scientific and technical outcomes. Interested parties should become familiar with the EMSL facility and capabilities (http://www.emsl.pnl.gov/) and contact the EMSL Director for further information. An external review committee for the biology grand challenge will select concepts for further development. EMSL will host a workshop(s) for the successful concept team(s) to develop the scientific plan, strategy, and resource requirements for the grand challenge, which will be used in the final selection process. The EMSL staff will assist successful teams in implementing their program.

In Search of Complexity: Bioinformatics and Molecular Tools in the Search for Protein Recoding

Norma Wills, Barry Moore, Andy Hammer, Mike Howard, Clark Henderson, Jason Simmons, John Atkins & Ray Gesteland
Department of Human Genetics, University of Utah, Salt Lake City, Utah, USA

Estimates of the complexity of the human genome are converging to somewhere around 30,000 genes. This is only 5 times more than the number of genes in the simplest eukaryote – the single celled yeast, and less that twice the number of genes found in the worm. How do we account for the much greater complexity of the human? We must look to the proteins. It is well known that the number of unique proteins expressed by the genome is much greater than the number of genes. A number of mechanisms add to the complexity of the proteome. Some of these mechanisms such as alternative splicing and post-translational modification are well understood. Other sources of complexity are just beginning to be understood such as, programmed ribosomal frameshifting, transcriptional slippage, RNA editing and ribosomal bypassing. In the field of proteomics the study of peptides generated from full length proteins by protease cleavage remains the standard protocol. Peptides are readily removed from 2-D gels which have traditionally been proteomic's separation tool of choice, and they are much more efficiently separated by 2-D liquid chromatography which has been increasingly popular in recent years as a tool for proteome separation. However with all the benefits of peptide proteomics, valuable information is lost when full length proteins aren't considered.

Separation of very complex mixtures of full length proteins into fractions for analysis by mass spectrometry is exceptionally challenging. Abundant proteins are routinely found throughout the separated fractions, masking the presence of less abundant proteins, or those for which ionization is difficult. Tagging full length proteins and expressing them under their native promoter, or with over expression, is an alternative for preparing high quality protein samples for mass spec analysis. However, tagging, purifying and mass analyzing all known and predicted proteins from a genome would be monumental task. Here bioinformatics can inform our decisions about which proteins or predicted proteins are most likely to show non-standard decoding.

Here we discuss work that involves extensive searching of the human genome for telltale signs of recoding. We use Perl and Java (along with the biological counterparts Bioperl, and Biojava) to search predicted gene regions for frameshifting motifs, transcriptional slippage sites, RNA secondary structure, unusual ORF architecture, and cross species conservation to find potential targets cases of recoding. We use β-galactosidase and dual luciferase fusions to assay for frameshifting, readthrough, and transcriptional slippage. We employ affinity-tagging protein purification to isolate recoded protein products, and electrospray ionization mass spectrometry to identify the full length protein mass. This coupled with site specific mutagenesis allows us to tease apart the molecular details of known and predicted translational recoding events. As we

learn more about these details of recoding we will be better able to apply that knowledge to de novo prediction of recoding sites in raw genomic sequence.

Appendix 1: Attendees List

Attendees list as of February 9, 2004.

Michael Adams
University of Georgia
adams@bmb.uga.edu

Jon Amster
University of Georgia
jamster@uga.edu

Jeff Amthor
U.S. Department of Energy
jeff.amthor@science.doe.gov

Gordon Anderson
Pacific Northwest Laboratory
gordon.anderson@pnl.gov

George Andrews Jr.
NSWCDD, B55
andrewsga@nswc.navy.mil

Adam Arkin
Lawrence Berkeley National Laboratory
aparkin@lbl.gov

Daniel Arp
Oregon State University
arpd@science.oregonstate.edu

Steve Ashby
Lawrence Livermore National Laboratory
garrigan2@llnl.gov

Gyorgy Babnigg
Argonne National Laboratory
gbabnigg@anl.gov

Holly Baden-Tillson
Institute for Biological Energy Alternatives
HBaden-Tillson@bioenergyalts.org

Gang Bao
Georgia Institute of Technology/Emory University
gang.bao@bme.gatech.edu

John Battista
Louisiana State University
jbattis@lsu.edu

Jeffrey Bernstein
University of California, Los Angeles
jbern1@ucla.edu

Harvey Bolton
Pacific Northwest National Laboratory
harvey.bolton@pnl.gov

Daniel Bond
University of Massachusetts
dbond@nre.umass.edu

Jennifer Bownas
Oak Ridge National Laboratory
bownasjl@ornl.gov

Tim Boyle
Sandia National Laboratories
tjboyle@Sandia.gov

Andrew Bradbury
Los Alamos National Laboratory
amb@lanl.gov

Elbert Branscomb
Lawrence Livermore National Laboratory
branscomb1@llnl.gov

Fred Brockman
Pacific Northwest National Laboratory
fred.brockman@pnl.gov

Michelle Buchanan
Oak Ridge National Laboratory
buchananmv@ornl.gov

Jessica Butler
University of Massachusetts, Amherst
jbutler@microbio.umass.edu

David Case
The Scripps Research Institute
case@scripps.edu

Denise Casey
Oak Ridge National Laboratory
caseydk@ornl.gov

Jamie Cate
University of California, Berkeley
jcate@lbl.gov

Michael Chandler
CNRS Labo. Microbio. et Genetique Moleculaire
Mike@ibcg.biotoul.fr

Parag Chitnis
National Science Foundation
pchitnis@nsf.gov

Linda Chrisey
Office of Naval Research
chrisel@onr.navy.mil

Ray-Yuan Chuang
Institute for Biological Energy Alternatives
rchuang@bioenergyalts.org

George Church
Harvard
g1m1c1@arep.med.harvard.edu

Stacy Ciufo
University of Massachusetts
sciufo@microbio.umass.edu

Dean Cole
U.S. Department of Energy/OBER
dean.cole@science.doe.gov

Frank Collart
Argonne National Laboratory
fcollart@anl.gov

James Collins
Boston University
jcollins@bu.edu

Maddalena Coppi
University of Massachusetts, Amherst
mcoppi@microbio.umass.edu

Jennifer Couch
National Cancer Institute
couchj@mail.nih.gov

Robert Coyne
National Science Foundation
rcoyne@nsf.gov

Robert DeBoy
The Institute for Genomic Research
rdeboy@tigr.org

Mitchel Doktycz
Oak Ridge National Laboratory
doktyczmj@ornl.gov

Timothy Donohue
University of Wisconsin, Madison
tdonohue@bact.wisc.edu

Kenneth Downing
Lawrence Berkeley National Laboratory
khdowning@lbl.gov

Sharon Doyle
DOE Joint Genome Institute
sadoyle@lbl.gov

Daniel Drell
U.S. Department of Energy
daniel.drell@science.doe.gov

Inna Dubchak
Lawrence Berkeley National Laboratory
ildubchak@lbl.gov

Susan Ehrlich
Arizona Court of Appeals
sehrlich@courts.sp.state.az.us

Leland Ellis
Department of Homeland Security
leland.ellis@dhs.gov

Peg Folta
Lawrence Livermore National Laboratory
folta2@llnl.gov

Jim Fredrickson
Pacific Northwest National Laboratory
jim.fredrickson@pnl.gov

Robert Friedman
Institute for Biological Energy Alternatives
rfriedman@ tcag.org

Teresa Fryberger
U.S. Department of Energy
teresa.fryberger@science.doe.gov

Daniel Gallahan
National Cancer Institute
dg13w@nih.gov

Timothy Gardner
Boston University
tgardner@bu.edu

Nataie Gassman
University of California, Los Angeles
ngassman@chem.ucla.edu

Al Geist
Oak Ridge National Laboratory
gst@ornl.gov

Raymond Gesteland
University of Utah
ray.gesteland@genetics.utah.edu

Carol Giometti
Argonne National Laboratory
csgiometti@anl.gov

Stephen Giovannoni
Oregon State University
steve.giovannoni@orst.edu

John Glass
Institute for Biological Energy Alternatives
john.glass@bioenergyalts.org

Andrey Gorin
Oak Ridge National Laboratory
agor@ornl.gov

Deborah Gracio
Pacific Northwest National Laboratory
debbie.gracio@pnl.gov

David Haaland
Sandia National Laboratories
dmhaala@sandia.gov

Murray Hackett
University of Washington
mhackett@u.washington.edu

Ioana Hance
The Institute for Genomic Research
ihance@tigr.org

Terry Hazen
Lawrence Berkeley National Laboratory
tchazen@lbl.gov

Grant Heffelfinger
Sandia National Laboratories
gsheffe@sandia.gov

Osnat Herzberg
University of Maryland Biotech Institute
osnat@carb.nist.gov

Robert Hettich
Oak Ridge National Laboratory
hettichrl@ornl.gov

Peter Highnam
National Institutes of Health/NCRR
highnam@NIH.gov

Colin Hill
Gene Network Sciences
colin@gnsbiotech.com

Roland F. Hirsch
U.S. Department of Energy
roland.hirsch@science.doe.gov

Lynette Hirschman
MITRE
lynette@mitre.org

Hoi-Ying Holman
Lawrence Berkeley National Laboratory
hyholman@lbl.gov

Norman Hommes
Oregon State University
hommesn@science.oregonstate.edu

Brian Hooker
Pacific Northwest National Laboratory
brian.hooker@pnl.gov

John Houghton
U.S. Department of Energy
john.houghton@science.doe.gov

Peter Hoyt
Oak Ridge National Laboratory
hoytpr@ornl.gov

Greg Hurst
Oak Ridge National Laboratory
hurstgb@ornl.gov

Thomas Huser
Lawrence Livermore National Laboratory
huser1@llnl.gov

Janet Jacobsen
Lawrence Berkeley National Laboratory
jsjacobsen@lbl.gov

Eric Jakobsson
University of Illinois, Urbana-Champaign
jake@ncsa.uiuc.edu

Grant Jensen
California Institute of Technology
jensen@caltech.edu

Gary Johnson
U.S. Department of Energy
garyj@er.doe.gov

Samuel Kaplan
University of Texas Medical School, Houston
Samuel.Kaplan@uth.tmc.edu

Arthur Katz
U.S. Department of Energy
arthur.katz@science.doe.gov

Jay Keasling
Lawrence Berkeley National Laboratory
keasling@socrates.berkeley.edu

Martin Keller
Diversa Corporation
mkeller@diversa.com

Stephen Kennel
Oak Ridge National Laboratory
kennelsj@ornl.gov

Stephen Kent
University of Chicago
skent@uchicago.edu

David Kirchman
University of Delaware
kirchman@cms.udel.edu

Peter Kirchner
Oak Ridge National Laboratory
kirchnep@mail.nih.gov

Joel Klappenbach
Michigan State University
klappenb@msu.edu

Eugene Kolker
BIATECH
ekolker@biatech.org

Julia Krushkal
University of Tennessee, Memphis
jkrushka@utmem.edu

Henrietta Kulaga
IPTO, DARPA
hkulaga@snap.org

Mike Kuperberg
U.S. Department of Energy
michael.kuperberg@science.doe.gov

Vladimir Kuznetsov
National Institutes of Health
kuznetsv@mail.nih.gov

Todd Lane
Sandia National Laboratories
twlane@sandia.gov

Frank Larimer
Oak Ridge National Laboratory
larimerfw@ornl.gov

Kim Lewis
Northeastern Uniersity
k.lewis@neu.edu

Huilin Li
Brookhaven National Laboratory
hli@bnl.gov

James Liao
University of California, Los Angeles
liaoj@ucla.edu

Mary Lipton
Pacific Northwest National Laboratory
Mary.Lipton@pnl.gov

John Logsdon
University of Iowa
john-logsdon@uiowa.edu

Derek Lovley
University of Massachusetts
dlovley@microbio.umass.edu

Suneeta Mandava
Argonne National Laboratory
smandava@anl.gov

Reinhold Mann
Oak Ridge National Laboratory
mannrc@ornl.gov

Betty Mansfield
Oak Ridge National Laboratory
mansfieldbk@ornl.gov

Costas Maranas
Pennsylvania State University
costas@psu.edu

Uljana Mayer-Cumblidge
Pacific Northwest National Laboratory
Uljana.Mayer-Cumblidge@pnl.gov

Harley McAdams
Stanford University
hmcadams@stanford.edu

Lee Ann McCue
Wadsworth Center
mccue@wadsworth.org

Barbara Methe
The Institute for Genomic Research
bmethe@tigr.org

Noelle Metting
U.S. Department of Energy
noelle.metting@science.doe.gov

George Michaels
Pacific Northwest National Laboratory
tara.hoyem@pnl.gov

Marissa Mills
Oak Ridge National Laboratory
millsmd@ornl.gov

Julie Mitchell
University of Wisconsin, Madison
mitchell@math.wisc.edu

Emmanuel Mongodin
The Institute for Genomic Research
emongodin@tigr.org

Jennifer Morrell
Oak Ridge National Laboratory
morrelljl1@ornl.gov

Sue Morss
Argonne National Laboratory
smorss@anl.gov

Shreedhar Natarajan
University of Illinois, Urbana Champaign
natarajn@ncsa.uiuc.edu

Karen Nelson
The Institute for Genomic Research
kenelson@tigr.org

Philippe Normand
CNRS UMR Ecologie Microbienne
normand@biomserv.univ-lyon1.fr

Kim Nylander
Oak Ridge National Laboratory
nylanderk@ornl.gov

Yasuhiro Oda
University of Iowa
yasuhiro-oda@uiowa.edu

Susan E. Old
National Institutes of Health
olds@nhlbi.nih.gov

Carl Oliver
U.S. Department of Energy
ed.oliver@science.doe.gov

Paula Olsiewski
Alfred P. Sloan Foundation
olsiewski@sloan.org

Himadri Pakrasi
Washington University
Pakrasi@wustl.edu

Brian Palenik
University of California, San Diego
bpalenik@ucsd.edu

Dina Paltoo
National Institutes of Health
paltood@mail.nih.gov

Ari Patrinos
U.S. Department of Energy
ari.patrinos@science.doe.gov

Dale Pelletier
Oak Ridge National Laboratory
pelletierda@ornl.gov

Sam Purvine
BIATECH
spurvine@biatech.org

Pankaj Qasba
National Institutes of Health
qasba@nhlbi.nih.gov

Gemma Reguera
University of Massachusetts
greguera@microbio.umass.edu

Karin Remington
Institute for Biological Energy Alternatives
kremington@tcag.org

Haluk Resat
Pacific Northwest National Laboratory
haluk.resat@pnl.gov

Monica Riley
Marine Biological Laboratory
mriley@mbl.edu

Diane Rodi
Argonne National Laboratory
drodi@anl.gov

Margie Romine
Pacific Northwest National Lab
margie.romine@pnl.gov

Denise Russo
National Institutes of Health
drusso@mail.nih.gov

Andrey Rzhetsky
Columbia University
ar345@columbia.edu

Herbert Sauro
Keck Graduate Institute
hsauro@kgi.edu

Luis Sayavedra-Soto
Oregon State University
sayavedl@science.oregonstate.edu

Charlene Schaldach
Lawrence Livermore National Laboratory
schaldach1@llnl.gov

Denise Schmoyer
Oak Ridge National Laboratory
schmoyerdd@ornl.gov

David Schwartz
University of Wisconsin, Madison
dcschwartz@facstaff.wisc.edu

Bari Scott
SoundVision Productions
bariscot@aol.com

Salvatore Sechi
National Institutes of Health
Salvatore_Sechi@nih.gov

Margrethe Serres
Marine Biological Laboratory
mserres@mbl.edu

Lucy Shapiro
Stanford University
shapiro@stanford.edu

Rob Siegel
Pacific Northwest National Laboratory
robert.siegel@pnl.gov

Nancy Slater
Lawrence Berkeley National Laboratory
naslater@lbl.gov

Hamilton Smith
Institute for Biological Energy Alternatives
hsmith@bioenergyalts.org

Harold Smith
University of Maryland
smithh@umbi.umd.edu

Richard Smith
Pacific Northwest National Laboratory
rds@pnl.gov

Thomas Squier
Pacific Northwest National Laboratory
thomas.squier@pnl.gov

Marvin Stodolsky
U.S. Department of Energy
Marvin.Stodolsky@science.doe.gov

T.P. Straatsma
Pacific Northwest National Laboratory
tps@pnl.gov

Bill Studier
Brookhaven National Laboratory
studier@bnl.gov

Hong-wei Sun
National Institutes of Health
sunh1@mail.nih.gov

John Sutherland
Brookhaven National Laboratory
jcs@bnl.gov

F. Robert Tabita
Ohio State University
tabita.1@osu.edu

Chin-Hsien (Emily) Tai
National Institutes of Health
taic@pop.nci.nih.gov

Michael Teresinski
U.S. Department of Energy
michael.teresinski@science.doe.gov

David Thomassen
U.S. Department of Energy
david.thomassen@science.doe.gov

Mehrdad Tondravi
National Institutes of Health
mt270t@nih.gov

Clifford Unkefer
Los Alamos National Laboratory
cju@lanl.gov

Sanjay Vashee
Institute for Biological Energy Alternatives
svashee@tcag.org

J. Craig Venter
Institute for Biological Energy Alternatives
jcventer@tcag.org

Wim Vermaas
Arizona State University
wim@asu.edu

Michael Viola
U.S. Department of Energy
michael.viola@science.doe.gov

Lawrence Wackett
University of Minnesota
wackett@biosci.cbs.umn.edu

Judy Wall
University of Missouri, Columbia
wallj@missouri.edu

Xiufeng Wan
University of Missouri, Columbia
wanx@missouri.edu

Bruce Weaver
The Institute for Genomic Research
bweaver@tigr.org

Bobbie-Jo Webb-Robertson
Pacific Northwest National Laboratory
Bobbie-Jo.Webb-Robertson@pnl.gov

Xueming Wei
Oregon State University
weixue@science.oregonstate.edu

Bert Weinstein
Lawrence Livermore National Laboratory
weinstein2@llnl.gov

Shimon Weiss
University of California, Los Angeles
sweiss@chem.ucla.edu

Jean Weissenbach
GENOSCOPE
jsbach@genoscope.cns.fr

Owen White
The Institute for Genomic Research
owhite@tigr.org and jfowler@tigr.org

Julian Whitelegge
University of California, Los Angeles
jpw@chem.ucla.edu

C. John Whitmarsh
National Institutes of Health
whitmarj@nigms.nih.gov

Sharon Wiback
Genomatica, Inc
swiback@genomatica.com

H. Steven Wiley
Pacific Northwest National Laboratory
steven.wiley@pnl.gov

Stan Wullschleger
Oak Ridge National Laboratory
wullschlegsd@ornl.gov

X. Sunney Xie
Harvard University
xie@chemistry.harvard.edu

Qing Xu
Institute for Biological Energy Alternatives
qxu@tcag.org

Ying Xu
University of Georgia
xyn@bmb.uga.edu

Jane Ye
National Heart, Lung, and Blood Institute
yej@nhlbi.nih.gov

Shibu Yooseph
Institute for Biological Energy Alternatives
shibu.yooseph@bioenergyalts.org

Zhaoduo Zhang
Sandia National Laboratories
zzhang@sandia.gov

Jizhong Zhou
Oak Ridge National Laboratory
zhouj@ornl.gov

Appendix 2: Web Sites

Genomics:GTL Web Sites

- Home page: http://doegenomestolife.org
- Genomics:GTL Roadmap, April 2001: http://doegenomestolife.org/roadmap/GTLcomplete.pdf
- Program Overview, January 2003: http://doegenomestolife.org/pubs/overview_screen.pdf
- Genomics: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
- Workshop Reports on Computing, Technologies, and Facilities: http://doegenomestolife.org/pubs.shtml
- Payoffs for the Nation: Energy Security and Global Climate Change; Bioremediation: http://doegenomestolife.org/pubs.shtml#payoffs
- Funded Projects: http://www.doegenomestolife.org/research/index.shtml
- Publications and Presentations: http://doegenomestolife.org/pubs.shtml
- Image Gallery: http://www.ornl.gov/hgmis/graphics/slides/images3.shtml

Complementary Web Sites

- DOE Microbial Genome Program: http://www.ornl.gov/microbialgenomes/
- Microbial Genomics Gateway: http://microbialgenome.org/
- Human Genome Project Information: http://www.ornl.gov/hgmis/

Author Index

Primary authors are indicated in bold.

D

E

F

G

H

Primary authors are indicated in bold.

J

K

L

Primary authors are indicated in bold.

M

N

O

Primary authors are indicated in bold.

P

Q

R

S

Primary authors are indicated in bold.

T

U

V

W

Primary authors are indicated in bold.

X

Y

Z

Primary authors are indicated in bold.

Institution Index

N

O

P

R

S

T

U

W

AGENDA

Genomics:GTL Contractor-Grantee Workshop II

February 29—March 2, 2004
Marriott Wardman Park, Washington, D.C.

Sunday, February 29, 2004

5:00–9:00 p.m. Registration and Poster Set Up

7:00–9:00 p.m. No Host Mixer and Poster Set Up

Monday, March 1, 2004

7:30–8:30	Continental Breakfast and Registration, Pre-Function Area
8:30–8:45	Welcome and Logistics—David Thomassen, BER, DOE Office of Science
8:45–9:30	GTL Project Update—Derek Lovley, Univ. of Mass., Amherst (See page 31)
9:30–10:00	BREAK
10:00–10:45	GTL Project Update, Craig Venter, Institute for Biological Energy Alternatives (See page 51)
10:45–11:30	GTL Project Update, Adam Arkin, LBNL (See page 3)
11:30–12:00	The Production Genomics Facility as a User Facility–The Community Sequencing Project, Eddy Rubin, JGI
12:00–1:30	Lunch On Your Own
1:30–2:00	GTL Program Overview and Goals, Ari Patrinos, Director for Biological and Environmental Research, DOE Office of Science
2:00–2:45	Computational and Informatics Challenges for the GTL Program, Ed Uberbacher, ORNL
2:45–3:00	BREAK
3:00–4:30	Breakout Discussion Groups
4:30–8:00	Poster Session Poster boards are numbered; poster placement corresponds to abstract number.

Tuesday, March 2, 2004

7:45–8:30	Continental Breakfast, Pre-Function Area
8:30 - 8:45	Introduction to Scientific Computing at DOE—Ed Oliver, Director for Advanced Scientific Computing Research, DOE Office of Science
8:45–9:15	GTL Computing Overview—Gary Johnson, Office of Advanced Scientific Computing, DOE Office of Science
9:15–10:00	GTL Project Update—Grant Heffelfinger, SNL (See page 20)
10:00–10:15	BREAK
10:15–11:00	GTL Project Update—George Church, Harvard University (See page 1)
11:00–11:45	GTL Project Update—Michelle Buchanan and Steve Wiley, ORNL/PNNL (See page 13)
11:45–1:00	Lunch On Your Own
1:00–1:30	Opportunities for Interaction with the Biotechnology Industry Organization (BIO)—Brent Erickson, Vice President, Industrial and Environmental Section, BIO
1:30–2:15	Town Meeting on Data Exchange and Data Standards—Adam Arkin (discussion leader), LBNL
2:15–3:00	GTL Project Update—Jim Fredrickson, PNNL (See page 38)
3:00	Closing Comments and Adjourn—David Thomassen

Breakout Sessions

1. Ecogenomics and Microbial Community Function.
   Discussion co-chairs: Eddy Rubin, JGI and Craig Venter, IBEA
2. Proteins and Protein Tags.
   Discussion co-chairs: Michelle Buchanan, ORNL and Andrew Bradbury, LANL
3. Imaging.
   Discussion co-chairs: Sunney Xie, Harvard and Ken Downing, LBNL
4. Microbial Cultivation Technologies.
   Discussion co-chairs: Brian Davison, ORNL and Yuri Gorby, PNNL

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