DOE/SC-0096

DOE GENOMICS:GTL
SYSTEMS BIOLOGY
FOR ENERGY AND
ENVIRONMENT

OFFICE OF SCIENCE
U.S. DEPARTMENT OF ENERGY

Joint Meeting

Genomics:GTL Contractor-Grantee Workshop IV

and

Metabolic Engineering Working Group Inter-Agency Conference on Metabolic Engineering 2006

North Bethesda, Maryland
February 12–15, 2006

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

http://www.doegenomestolife.org

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

Welcome to GTL-MEWG Workshop

Welcome to the 2006 joint meeting of the fourth Genomics:GTL Contractor-Grantee Workshop and the sixth Metabolic Engineering Working Group Inter-Agency Conference. The vision and scope of the Genomics:GTL program continue to expand and encompass research and technology issues from diverse scientific disciplines, attracting broad interest and support from researchers at universities, DOE national laboratories, and industry. Metabolic engineering's vision is the targeted and purposeful alteration of metabolic pathways to improve the understanding and use of cellular pathways for chemical transformation, energy transduction, and supramolecular assembly. These two programs have much complementarity in both vision and technological approaches, as reflected in this joint workshop.

GTL's challenge to the scientific community remains the further development and use of a broad array of innovative technologies and computational tools to systematically leverage the knowledge and capabilities brought to us by DNA sequencing projects. The goal is to seek a broad and predictive understanding of the functioning and control of complex systems —individual microbes, microbial communities, and plants. GTL's prominent position at the interface of the physical, computational, and biological sciences is both a strength and a challenge. Microbes remain GTL's principal biological focus. In the complex "simplicity" of microbes, we find capabilities needed by DOE and the nation for clean and secure energy, cleanup of environmental contamination, and sequestration of atmospheric carbon dioxide that contributes to global warming. An ongoing challenge for the entire GTL community is to demonstrate that the fundamental science conducted in each of your research projects brings us a step closer to biology-based solutions for these important national energy and environmental needs.

This year marks an important milestone for GTL with release of the roadmap that will help guide and justify the GTL program to a broad audience of scientists, policymakers, and the public. This important document was developed through a process of broad community participation that included many of you. It traces the path from DOE mission science through systems microbiology to the promise of emerging technologies, integrated computing, and a new research infrastructure. It describes opportunities, research strategies, and solutions related to this new science as applied to microbes and the complexities of mission problems.

To make GTL science and biological research broadly tractable, timely, and affordable, GTL will institute four user facilities to deliver economies of scale and enhance performance. These facilities will provide advanced technologies and state-of-the art computing needed to better understand genomic capability, cellular responses, regulation, and community behaviors in any environment. Another important step forward for GTL is the solicitation of applications for development of the Facility for Production and Characterization of Proteins and Molecular Tags, first of the four planned.

This year's GTL-MEWG workshop provides 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. We look forward to a stimulating and productive meeting and offer our sincere thanks to the organizers and to you, the scientists, whose vision and efforts will help us all to realize the promise of this exciting research program.

Patrinos

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

Michael Strayer
Associate Director of Science for
Advanced Scientific Computing
Research
Office of Science
U.S. Department of Energy
michael.strayer@science.doe.gov

Fred Heineken

Fred Heineken
Chair of the Interagency Metabolic
Engineering Working Group
National Science Foundation
fheineke@nsf.gov

Contents

* Presenting author

* Presenting author

Protein Production and Characterization 29

* Presenting author

Molecular Interactions 40

* Presenting author

* Presenting author

Milestone 2

* Presenting author

* Presenting author

* Presenting author

* Presenting author

* Presenting author

* Presenting author

* Presenting author

Poster Page

Milestone 3

Computing Infrastructure and Education

* Presenting author

Workshop Abstracts

This is the first Genomics:GTL workshop for principal investigators since the release of *Genomics:GTL Roadmap: Systems Biology for Energy and Environment* in October 2005 (www.doegenomestolife.org). Abstracts and posters for this workshop are organized around the GTL goal and milestones shown below and in the roadmap on pp. 42–55. Many of the research projects are essentially pilots or proof-of-principle studies for systems biology, technology and methods development, computing, and facilities.

Most computing abstracts, instead of comprising a separate category, are associated with appropriate experimental topics. Overarching computing infrastructure and education abstracts are under Milestone 3, GTL Computational Biology Environment.

Abstracts of the Metabolic Engineering Working Group (MEWG), an interagency approach to understanding and using metabolic processes, are identified as such and intermixed with GTL abstracts in relevant program categories.

Genomics:GTL Overarching Scientific Goal

Achieve a predictive, systems-level understanding of biological systems to help enable biobased solutions to DOE mission challenges.

Science and Technology Milestones

Milestone 1: The Code—Understand Gene Structure and Functional Potential of Plants and Microbes and Their Communities

- Organism Sequencing, Annotation, and Comparative Genomics
- Microbial-Community Sequencing
- Protein Production and Characterization
- Molecular Interactions

Milestone 2: The Response—Understand Function, Regulation, and Dynamics in Plants and Microbes and Their Communities

- Omics: Systems Measurements of Plants, Microbes and Communities
- Metabolic Network Experimentation and Modeling
- Regulatory Processes

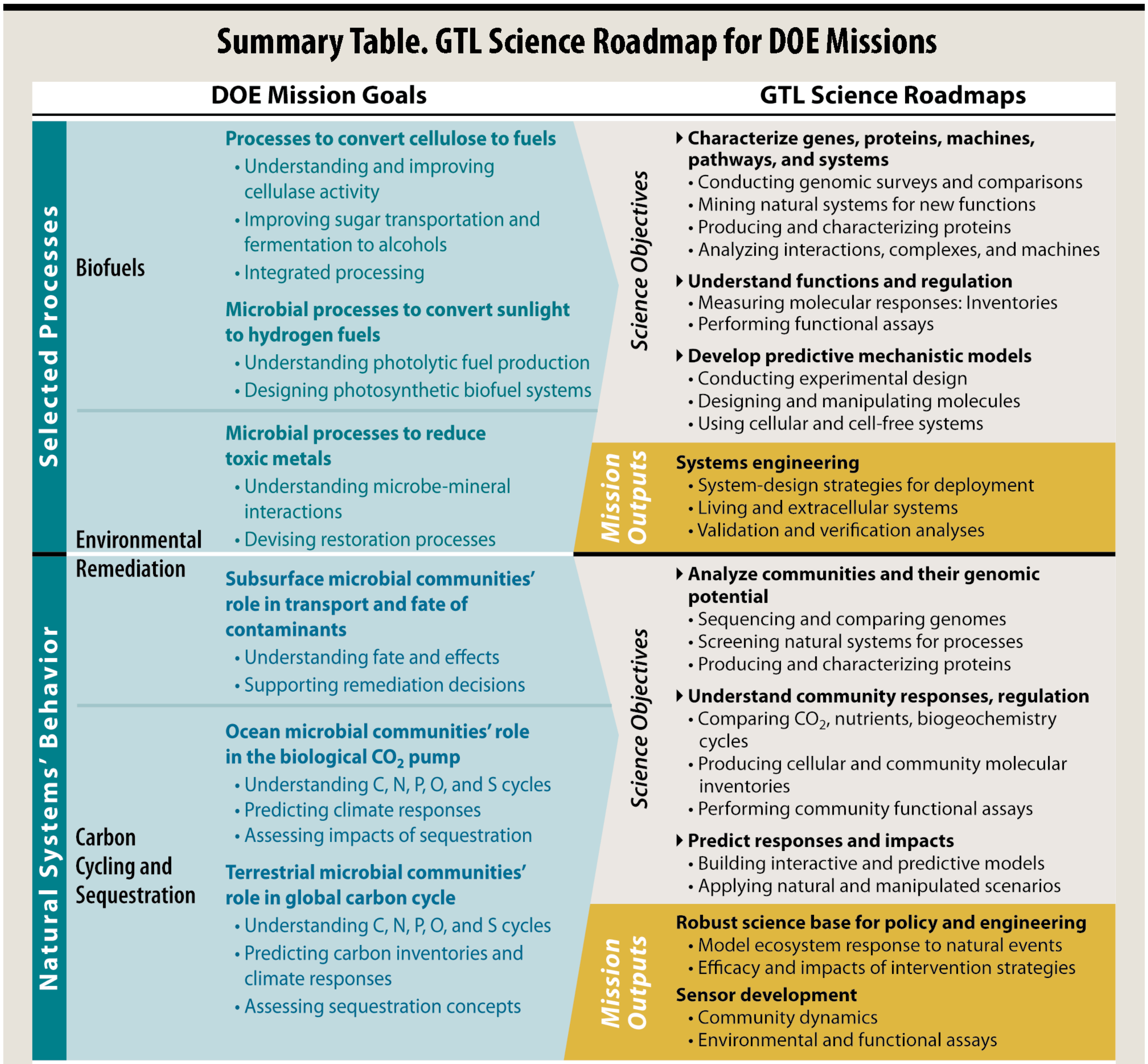
Milestone 3: GTL Computational Biology Environment

- Computing Infrastructure and Education

Communication

Ethical, Legal, and Societal Issues

The following table is a simple summation of how GTL science and DOE missions align (GTL Roadmap p. 40).



Summary Table. GTL Science Roadmap for DOE Missions

DOE Mission Goals

Selected Processes

Biofuels

Processes to convert cellulose to fuels
- Understanding and improving cellulase activity
- Improving sugar transportation and fermentation to alcohols
- Integrated processing

Microbial processes to convert sunlight to hydrogen fuels
- Understanding photolytic fuel production
- Designing photosynthetic biofuel systems

Environmental Remediation

Microbial processes to reduce toxic metals
- Understanding microbe-mineral interactions
- Devising restoration processes

Natural Systems' Behavior

Subsurface microbial communities' role in transport and fate of contaminants
- Understanding fate and effects
- Supporting remediation decisions

Carbon Cycling and Sequestration

Ocean microbial communities' role in the biological CO_2 pump
- Understanding C, N, P, O, and S cycles
- Predicting climate responses
- Assessing impacts of sequestration

Terrestrial microbial communities' role in global carbon cycle
- Understanding C, N, P, O, and S cycles
- Predicting carbon inventories and climate responses
- Assessing sequestration concepts

GTL Science Roadmaps

Science Objectives

Characterize genes, proteins, machines, pathways, and systems
- Conducting genomic surveys and comparisons
- Mining natural systems for new functions
- Producing and characterizing proteins
- Analyzing interactions, complexes, and machines

Understand functions and regulation
- Measuring molecular responses: Inventories
- Performing functional assays

Develop predictive mechanistic models
- Conducting experimental design
- Designing and manipulating molecules
- Using cellular and cell-free systems

Mission Outputs

Systems engineering
- System-design strategies for deployment
- Living and extracellular systems
- Validation and verification analyses

Science Objectives

Analyze communities and their genomic potential
- Sequencing and comparing genomes
- Screening natural systems for processes
- Producing and characterizing proteins

Understand community responses, regulation
- Comparing CO_2, nutrients, biogeochemistry cycles
- Producing cellular and community molecular inventories
- Performing community functional assays

Predict responses and impacts
- Building interactive and predictive models
- Applying natural and manipulated scenarios

Mission Outputs

Robust science base for policy and engineering
- Model ecosystem response to natural events
- Efficacy and impacts of intervention strategies

Sensor development
- Community dynamics
- Environmental and functional assays

A capsule summary of systems being studied, mission goals that drive the analysis, generalized science roadmaps, and outputs to DOE missions. To elucidate design principles, each of these goals entails the examination of thousands of natural primary and ancillary pathways, variants, and functions, as well as large numbers of experimental mutations.

Milestone 1

The Code—Understand Gene Structure and Functional Potential of Plants and Microbes and Their Communities

Section 1

Organism Sequencing, Annotation, and Comparative Genomics

1

U.S. DOE Joint Genome Institute Microbial Sequencing: Genomes to Life Projects

David Bruce[1]* (dbruce@lanl.gov), Tom Brettin[1], Patrick Chain[3], Cliff Han[1], Loren Hauser[5], Nikos Kyrpides[2], Miriam Land[5], Alla Lapidus[2], Frank Larimer[5], Jeremy Schmutz[4], Paul Gilna[1], Eddy Rubin[2], and Paul Richardson[2]

[1]JGI-Los Alamos National Laboratory, Los Alamos, NM; [2]JGI-Production Genomics Facility and Lawrence Berkeley National Laboratory, Berkeley, CA; [3]JGI-Lawrence Livermore National Laboratory, Livermore, CA; [4]JGI-Stanford Human Genome Center, Palo Alto, CA; and [5]JGI-Oak Ridge National Laboratory, Oak Ridge, TN

The US DOE Joint Genome Institute (JGI) sequences microbial and metagenomic projects through three main programs: DOE Microbial Genome Program (MGP), JGI Community Sequencing Program (CSP) and DOE Genomes to Life Program (GTL). The principle goal of the MGP is to fund sequencing projects related to DOE interests, the principle goal of the CSP is to fund sequencing projects from a broad range of disciplines that may not be covered in the MGP, and the principle goal of the GTL sequencing projects is to fund sequencing projects in direct support of the GTL program. The JGI is responsible for sequencing, assembling, annotating microbial genomes, and publishing sequence and annotation in GenBank and the DOE JGI Integrated Microbial Genomics web based system. The JGI has sequenced nearly 250 microbes and metagenomic samples to draft quality and completely finished over 120 microbes. Most microbial projects are targeted for finishing. The overall capacity is now approximately 100-125 microbial projects per year through draft sequencing and finishing. Virtually all microbial projects are sequenced by the whole genome shotgun method. To being the sequencing process, the Library group randomly shears the purified DNA under different conditions and selects for three size populations. Fragments are end repaired and selected for inserts in the range of 3kb, 8kb, and 40kb. These are cloned into different vector systems and checked for quality by PCR or sequencing. The libraries are sequenced by the Production group to approximately 8.5X coverage. The resulting reads are trimmed for vector sequences and assembled. The assembly is quality checked, automatically annotated by the Annotation group, and released to the collaborating PI as the initial Quality Draft assembly. For finishing, the draft assembly is assigned to a Finishing group. The Finishing group closes all sequence gaps, resolves all repeat discrepancies, and improves all low quality regions. The final assembly is then passed to the Quality Assurance group to assess the integrity and overall quality of the genome sequence. The finished sequence then receives a final annotation and this package is used as the basis for analysis and publication in GenBank and the DOE JGI Integrated Microbial Genomics web based system. The JGI is made up of affiliates from a number of national laboratories including Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, Los Alamos National Laboratory, Oak Ridge National Laboratory, and the Stanford Human Genome Center.

* Presenting author

2

High Throughput Genome Annotation for U.S. DOE Joint Genome Institute Microbial Genomes

Miriam Land* (landml@ornl.gov), Loren Hauser, Phil LoCascio, Gwo-Liang Chen, Denise Schmoyer, and Frank Larimer

Oak Ridge National Laboratory, Oak Ridge TN

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

The U.S. DOE Joint Genome Institute (JGI) performs high-throughput sequencing and annotation of microbial genomes through the DOE Microbial Genome Program (MGP). The world-wide rate of sequencing is resulting in a rapid expansion of microbial genomic data, which requires the development of comprehensive automated tools to provide in-depth annotation which can keep pace with the expanding microbial dataset. We have and continue to develop tools for genome analysis that provide automated, regularly updated, comprehensive annotation of microbial genomes using consistent methodology for gene calling and feature recognition. We have developed and continue to improve a genome annotation pipeline. The pipeline includes gene calls, multiple database searches, prediction of RNAs, and other annotation tools as they become available for a diverse and automated annotation.

Comprehensive representation of microbial genomes requires deeper annotation of structural features, including operon and regulon organization, promoter and ribosome binding site recognition, miscellaneous RNAs, and other functional elements. Linkage and integration of the gene/protein/function catalog to phylogenomic, structural, proteomic, transcriptional, and metabolic profiles are being developed. The expanding set of microbial genomes comprises an extensive resource for comparative genomes: new tools continue to be developed for rapid exploration of gene and operon phylogeny, regulatory networking, and functional proteomics.

A major continuing activity involves the public release of the data. Each genome is supported with a web site of the automated annotation, the data are submitted to GenBank for broader release and the data are prepared for the JGI's Integrated Microbial Genomes (IMG) database. The IMG resource is updated quarterly, in addition to the continuous addition of new genomes from JGI. 50-100 new projects will be initiated annually by JGI that require annotation. The deep sequencing of specific genera as well as specialized (physiological and phylogenetic) groups requires new views and analytical schemes.

The JGI is made up of affiliates from a number of national laboratories including Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, Los Alamos National Laboratory, Oak Ridge National Laboratory, and the Stanford Human Genome Center.

* Presenting author

3

Understanding Microbial Genomic Structures and Applications to Biological Pathway Inference

Z. Su[1], F. Mao[1], H. Wu[1], P. Dam[1], X. Chen[2], T. Jiang[2], V. Olman[1], B. Palenik[3], and **Ying Xu**[1]* (xyn@bmb.uga.edu)

[1]University of Georgia, Athens, GA; [2]University of California, Riverside, CA; and [3]Scripps Institution of Oceanography, University of California, San Diego, CA

The rapid increase in the number of sequenced microbial genomes provides unprecedented opportunities to computational biologists to decipher the genomic structures of these microbes through development and application of advanced comparative genome analysis tools. In this presentation, we describe a systematic study we have been carrying out on deciphering microbial genomic structures and linking the discovered genomic structures to prediction of metabolic pathways. This study consists of the following three main components: (a) deciphering microbial genomic structures and discovering new ones through development and application of advanced comparative genome analysis tools, (b) systematic study of relationships between microbial genomic structures and metabolic pathways through mapping all KEGG pathways to over 300 microbial genomes, and (c) application of the discovered relationships between genomic structures and pathways to prediction of biological pathways and networks.

A. Deciphering microbial genomic structures

We have recently developed a computer program JPOP[1,2] for operon structure predictions in both prokaryotic and archaea genomes. Testing on *E. coli.* data with experimental validation indicates that the program has an prediction accuracy about 80%. Since the publication of JPOP, a couple of operon prediction programs have been published including VIMSS[10] and Pathway Tools[11], reaching similar levels of prediction accuracy. Using these programs, we have made operon prediction for 300+ microbial genomes (all data are available upon request). This data set not only provides a rich source of information for our prediction of biological pathways and networks (see section C), but also facilitates investigation of higher level and less understood structures in microbial genomes. Through comparative genome analyses of 300+ microbial genomes, we have recently firmly established uber-operon, a concept introduced a few years ago by other authors, as a layer of genomic structures, which have direct implications to biological pathway predictions[3]. For example, we have demonstrated that a number of well studied metabolic pathways are made of (genes of) a small number of uber-operons (*versus* a large number of operons) [3]. In addition, we have established some interesting relationships between uber-operons and regulons, which have established a solid stepping stone for us to develop a computer program for regulon prediction in general *via* prediction of uber-operons. We have also recently developed an effective paradigm for predicting *cis* regulatory elements[4], through comparative analysis of closed related genomes, providing another important piece of information for regulon prediction. We expect that we will be able to develop the first computer program for regulon prediction in the very near future.

B. Systematic mapping of metabolic pathways to microbial genomes

The metabolic pathways of KEGG database provides a rich source of information, which can be directly mapped to individual genomes. However until very recently, there has not been an effective way for mapping KEGG pathways to genomes other than the simple minded approach through sequence similarity search. We have recently demonstrated that BLAST search or its variations/

* Presenting author

generalizations such as bi-direction best hit (BDBH) or COG search do not provide satisfactory mapping results[5] as virtually all these methods attempt to find orthologous gene relationship using sequence similarity information alone. We have recently developed a computer program P-MAP for mapping orthologous genes in the context of pathway mapping using both sequence similarity information and genomic structure information, having substantially improved the mapping accuracy of pathways. The basic idea of P-MAP pathway mapping is that it attempts to map genes of a pathway to their homologous genes in the target genome, under the condition that these mapped genes are grouped into a (small) number of operons. The limitation of the current P-MAP algorithm is that it assumes that a template pathway is given in a form that its individual components have genes assigned in the template genome, limiting direct applications of KEGG (template) pathways. We have recently generalized the framework of P-MAP, allowing mapping a generic pathway model (consisting of enzymes and enzymatic reasons rather than specific genes assigned to each enzyme) to a target genome, by mapping individual enzymes to genes that are grouped into a number of operons in the target genome[6]. Using this novel capability, we have mapped metabolic pathways of KEGG to 300+ microbial genomes (data are available upon request). A detailed analysis is currently under way, attempting to understand the general relationship between metabolic pathways and operon, uber-operon and regulon structures. We expect that this analysis will lead to new understanding about genomic structures, the organization and evolution of metabolic pathways, which is expected to be done within the next few weeks.

C. Pathway predictions through application of identified genomic structures:

As we understand now, genomic structures such as operons, uber-oprons and regulons and their detailed organizations provide significant amount of information about the component genes and even wiring diagrams of metabolic pathways. Using such information derived through prediction of operons, uber-operons and regulons, we have made a number of predictions of non-trivial biological networks, including phosphorus assimilation pathways[7], carbon fixation pathways[8], and nitrogen assimilation pathways[9] in *cyanobacteria* and a cross-talk network between nitrogen assimilation and photosynthesis[4], for which we for the first time proposed a detailed molecular mechanism how these two processes orchestrate with each other. These predictions have provided a number of new insights about these important biological processes. By extending these predictions, we are currently focused on prediction of a group of pathways relevant to carbon sequestration. Detailed results of these predicted pathway presentation will be presented at the GTL workshop.

Acknowledgement: This project is supported by the U.S. Department of Energy's Genomics:GTL Program under project "Carbon Sequestration in *Synechococcus sp*: From Molecular Machines to Hierarchical Modeling" (http://www.genomes-to-life.org).

References

1. Chen, X., Su, Z., Dam, P., Palenik, B., Xu, Y. and Jiang, T. (2004) "Operon prediction by comparative genomics: an application to the *Synechococcus sp.* WH8102 genome," *Nucleic Acids Res*, 32, 2147-2157.
2. Chen, X., Su, Z., Xu, Y. and Jiang, T. (2004) "Computational Prediction of Operons in Synechococcus sp. WH8102.," *Genome Inform Ser Workshop Genome Inform*, 15, 211-222 (best paper award).
3. Che, G. Li, F. Mao, H. Wu, and Ying Xu, "Detecting uber-operons in microbial genomes," submitted to *Proc Natl Acad Sci USA*, 2005.
4. Z. Su, F. Mao, V. Olman and Ying Xu, "Comparative genomics analyses of ntcA regulons in cyanobacteria: regulation of nitrogen assimilation and its coupling to photosynthesis," *Nucleic Acids Research*, vol 33(16): 5156 - 5171, 2005.
5. Mao, Z. Su, V. Olman, P. Dam, Z. Liu, Ying Xu, "Mapping of orthologous genes in the context of biological pathways: an application of integer programming," *Proc Natl Acad Sci USA*, 2005 (in press).

* Presenting author

6. Mao, W. Wu, Ying Xu, "Mapping of KEGG metabolic pathways to microbial genomes", submitted, 2005.
7. 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, *Genome Informatics* pp. 3 - 13, vol 14, Universal Academy Publishing, 2003.
8. P. Dam, Z. Su, V Olman, Ying Xu, *In silico* construction of the carbon fixation pathway in Synechococcus sp. WH8102, *Journal of Biological Systems*, vol. 12, pp.97-125, 2004.
9. Z. Su, P. Dam, F. Mao, V. Olman, I. Paulsen, B. Palenik and Ying Xu, Computational inference and experimental validation of nitrogen assimilation regulatory networks in cyanobacterium Synechococcus sp. WH8102, *Nucleic Acids Research* , 2005 (in press).
10. Price, M.N., et al., *A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic Acids Res*, 2005. 33(3): p. 880-92.
11. Karp, P.D., S. Paley, and P. Romero, *The Pathway Tools software.* Bioinformatics, 2002. 18 Suppl 1: p. S225-S232.

4 MEWG

The BioCyc Collection of 200 Pathway/Genome Databases and the MetaCyc Database of Metabolic Pathways and Enzymes

Peter D. Karp[1]* (pkarp@ai.sri.com), Christos Ouzounis[2], and Sue Rhee[3]

[1]SRI International, Menlo Park, CA; [2]European Bioinformatics Institute, Hinxton, UK; and [3]Carnegie Institution, Stanford, CA

The BioCyc Database Collection[1] is a set of 200 Pathway/Genome Databases (PGDBs) for most prokaryotic and eukaryotic organisms whose genomes have been completely sequenced to date. The BioCyc collection provides a unique resource for metabolic engineering and for global and comparative analyses of genomes and metabolic networks.

Each organism-specific PGDB within BioCyc contains the complete genome of the organism plus the following additional information inferred by the Pathway Tools[2] software:

- Predicted metabolic pathways as inferred from the MetaCyc[3] database
- Predicted genes to fill holes in the metabolic pathways (pathway holes are pathway steps for which no enzyme has been identified in the genome)
- Predicted operons for each bacterial PGDB
- Transport reactions inferred from the product descriptions of transport proteins by the Transport Inference Parser
- A metabolic overview diagram containing the metabolic enzymes, transport proteins, and membrane proteins of each organism is constructed automatically

The BioCyc collection can be accessed in several ways including interactive access via the BioCyc.org web site, bulk downloading in several formats including Systems Biology Markup Language (SBML) and BioPAX, and querying within SRI's BioWarehouse system for database integration. Most BioCyc PGDBs are freely and openly available to all.

* Presenting author

We seek scientists to adopt and curate individual PGDBs within the BioCyc collection. Only by harnessing the expertise of many scientists can we hope to produce biological databases that accurately capture the depth and breadth of biomedical knowledge. To adopt a database, send email to biocyc-support@ai.sri.com.

The Pathway Tools software that powers the BioCyc Web site provides powerful query and visualization operations for each BioCyc database. For example, the Omics viewer allows scientists to visualize combinations of gene expression, proteomics, and metabolomics data on the metabolic map of an organism (see http://biocyc.org/ov-expr.shtml). A genome browser permits interactive exploration of either a single genome, or of orthologous regions of multiple genomes. A newly developed set of comparative genomics tools supports many comparisons across the genomes and metabolic networks of the BioCyc collection. See http://biocyc.org/samples.shtml for an overview of BioCyc Web site functionality.

The MetaCyc database[3] describes experimentally elucidated metabolic pathways and enzymes as reported in the experimental literature. MetaCyc is both an online reference source on metabolic pathways and enzymes, and a solid foundation of experimentally proven pathways for use in computational pathway prediction. MetaCyc version 9.6 describes 690 pathways from more than 600 organisms. The 5500 biochemical reactions in MetaCyc reference 4800 chemical substrates, most of which contain chemical structure information. MetaCyc describes the properties of 3000 enzymes, such as their subunit structure, cofactors, activators, inhibitors, and in some cases their kinetic parameters. The information in MetaCyc was obtained from more than 8500 research articles, and emphasizes pathways and enzymes from microbes and plants.

References

1. P.D. Karp et al, "Expansion of the BioCyc Collection of Pathway/Genome Databases to 160 Genomes," *Nucleic Acids Research* 33:6083-9 2005.
2. P.D. Karp et al, "The Pathway Tools Software," *Bioinformatics* 18:S225-32 2002.
3. R. Caspi et al, "MetaCyc: A multiorganism database of metabolic pathways and enzymes," *Nucleic Acids Research* in press, 2006 Database issue.

5

Bioinformatic Methods Applied to Prediction of Bacterial Gene Function

Michelle Green[1]* (green@ai.sri.com), Balaji Srinivasan[2], Peter Karp[1], and **Harley McAdams**[2]

[1]SRI International, Menlo Park, CA and [2]Stanford University, Stanford, CA

We have applied two different bioinformatic methods to prediction of the function of *Caulobacter* gene products. First, we used the PathoLogic program to construct Pathway/Genome databases using the genome's annotation to predict the set of metabolic pathways present. PathoLogic determines the set of reactions composing those pathways from the list of enzymes in the organism. Enzymes in a genome are often missed or assigned a non-specific function (e.g., "thiolase family protein") during the initial annotation. These incomplete annotations result in "pathway holes" where the genome appears to lack the enzymes known to be needed to catalyze reactions in a pathway. Second, we combined gene co-conservation determined from 230 sequenced bacterial genomes with four

* Presenting author

types of functional genomic data to predict protein interaction probabilities and protein networks with greatly improved confidence compared to previous methods.

Increased coverage of PHFiller using genome context data. PHFiller, a previous algorithm developed by the Karp lab for filling pathway holes, utilized homology and pathway based evidence to determine the probability that a candidate enzyme filled a particular pathway hole (Green et al. 2004). Candidate enzymes for reaction R were identified by searching the organism's genome for homologs of a set of isozyme sequences from other organisms that catalyze reaction R. The algorithm does not identify candidates for pathway holes for which no isozyme sequences are available. Approximately 20% (44 pathway holes) of the remaining pathway holes in the CauloCyc Pathway/Genome Database (PGDB) are reactions for which such sequences are unavailable. We have increased the coverage of the PHFiller algorithm to include these reactions by incorporating genome context data into its search for candidate enzymes.

The protein complex ortholog method, a new source of genome context. In addition to the integration of phylogenetic profiles and gene neighborhood methods into the PHFiller algorithm, the Karp lab has developed an algorithm for identifying functionally associated gene pairs based on known protein complexes. If genes A and B in organism I are known to participate in a protein complex, then we infer that their orthologs, A' and B' in organism II, are functionally related. The EcoCyc PGDB describes 247 heterocomplexes, which yield almost 1400 protein pairs that participate in those complexes. The CauloCyc PGDB includes 158 protein pairs that are orthologs of the *E. coli* proteins pairs. Of these 158 pairs, 122 are annotated with the same COG functional category (Tatusov et al. 2003) and only 60 of these pairs have been identified with high confidence (confidence score greater than 0.7) in the STRING database (von Mering et al. 2003), indicating that our new method finds new functional relationships.

Integrated Protein Interaction Networks for 230 Microbes. The McAdams lab has combined four different types of functional genomic data to create high coverage protein interaction networks for 230 microbes. The integration algorithm naturally handles statistically dependent predictors and automatically corrects for differing noise levels and data corruption in different evidence sources. We find that many of the predictions in each integrated network hinge on moderate but consistent evidence from multiple sources rather than strong evidence from a single source, yielding novel biology which is missed if a single data source such as coexpression or coinheritance is used in isolation. In addition to statistical analysis and recapitulation of known biology, we demonstrate that these subtle interactions can discover new aspects of even well studied functional modules, such as the flagellar hierarchy and the cell division apparatus. This analysis has produced the largest collection of probabilistic protein interaction networks compiled to date, and the methods can be applied to any sequenced organism and any kind of experimental or computational technique which produces pairwise measures of protein interaction.

References

1. Green, M. L. and P. D. Karp (2004). "A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases." *BMC Bioinformatics* 5: 76.
2. Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin and D. A. Natale (2003). "The COG database: an updated version includes eukaryotes." *BMC Bioinformatics* 4: 41.
3. von Mering, C., M. Huynen, D. Jaeggi, S. Schmidt, P. Bork and B. Snel (2003). "STRING: a database of predicted functional associations between proteins." Nucleic Acids Res 31(1): 258-61.

6

Pipelining RDP Data to the Taxomatic and Linking to External Data Resources

S. H. Harrison[1], P. Saxman[1], T.G. Lilburn[2], J.R. Cole[1], and **G.M. Garrity**[1]* (garrity@msu.edu)

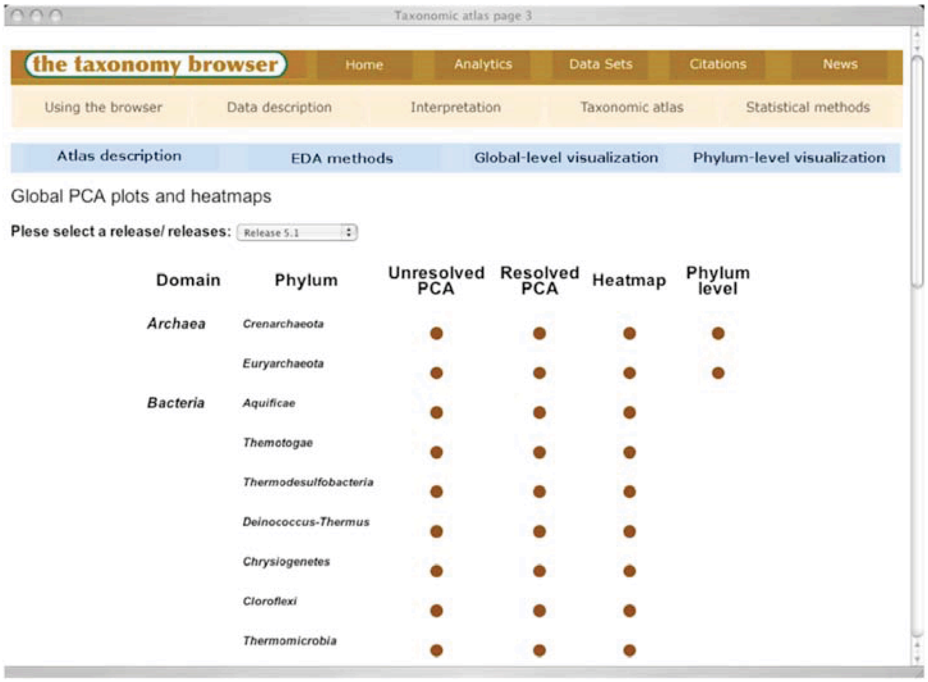
[1]Michigan State University, East Lansing, MI and [2]American Type Culture Collection, Manassas, VA

The taxonomic atlas represents an ongoing experiment in visualization of evolutionary relationships among the prokaryotes. Starting at a point of interest, the system allows users to move through a hierarchical classification, at different levels of taxonomic resolution, so that they may better gauge relationships based on a given gene, group of genes, or other quantitative signal that they might deem relevant. To demonstrate the potential of the methodology, we developed a data-driven atlas of taxonomic/phylogenetic heatmaps, based on a nomenclatural taxonomy that came to be known as the "Taxomatic". Since its inception, the prototype website has been moved to a production web server (http://taxoweb.mmg.msu.edu) where it is maintained and periodically updated. Methods were also developed to permit retrieval and side-by-side viewing of multiple interactive PCA plots from different releases of the underlying taxonomies to permit end-users to readily visualize changes that might otherwise go unrecognized using alternative techniques. For prototyping, we selected Insightful's StatServer as our platform for deploying our interactive graphics to the user community, but have reached the limits of this technology. In order to significantly reduce the amount of time needed to serve up the maps in the atlas, we have turned to AJAX-based technologies to develop a phylogenetic mapping service similar to the well-known geographical mapping service provided by Google. This provides server-side on-the-fly generation of image-mapped files that will ultimately link to key references in the literature, external sets of sequence and phenotypic data, and sources of viable cultures, when available. Together with browser-side JavaScript, these technologies can duplicate most (but not all) of the functionality available in Graphlets, but without the inherent client-side problems.

Using our tools, end-users can browse a taxonomic hierarchy, display, zoom, and re-center the view of a heatmap, highlight and display the names of the higher-level taxa fully visible on the current view, and adjust the taxonomic hierarchy to the show the taxon selected (clicked) on the heatmap by the user. At full zoom, it is possible to adjust the taxonomic hierarchy to identify the organism(s) selected (clicked) on the heatmap and link to external resources.

To build and maintain a carefully annotated and up-to-date reference classification we are connecting the "Taxomatic" to the RDP-II's database and tools, which will ensure timely gathering and rapid alignment of sequences. The models we build based on these alignments and provide via the "Taxomatic" must not only keep pace with the relevant sequence databases, but also must themselves be quickly and objectively built, organizing the data into an optimal or near optimal structure. This is being done using the SOSCC algorithm which was developed to automate this task and will play a central role in ensuring that the internal databases of both the "Taxomatic" and the RDP are relatively free of annotation errors and taxonomic anomalies that arise from the widely used nomenclatural model. The SOSCC code has been significantly revised and now supports both hierarchical and non-hierarchical classification techniques and varying levels of classification stringency. The code has been fully documented and has been packaged as an S-Plus library for distribution from the "Taxomatic" web site.

* Presenting author



Global PCA plots and heatmaps

Plese select a release/ releases: Release 5.1

Domain	Phylum	Unresolved PCA	Resolved PCA	Heatmap	Phylum level
Archaea	*Crenarchaeota*	●	●	●	●
	Euryarchaeota	●	●	●	●
Bacteria	*Aquificae*	●	●	●	
	Themotogae	●	●	●	
	Thermodesulfobacteria	●	●	●	
	Deinococcus-Thermus	●	●	●	
	Chrysiogenetes	●	●	●	
	Cloroflexi	●	●	●	
	Thermomicrobia	●	●	●	

Figure 1. A screenshot of the taxonomic atlas provided by the "Taxomatic".

Significant changes are occurring in the manner, style, and pace of scientific publications. Links to datasets and other online resources are becoming increasingly important to the scientific enterprise, yet traversing those links to relevant information and services while avoiding those that succumbed to "link-rot" (e.g. the notorious Error 404 – Link not found" problem) is an increasingly difficult challenge. While the ramifications of this problem have been discussed by Garrity and Lyons, implementation of a satisfactory solution has remained elusive. As an extension to this project we are investigating the feasibility of using interactive heatmaps and other graphics as navigational devices to link to a collection of persistently maintained "mini-monographs", which in turn provide persistent links to other available data, services, and information resources, specific to a given organism. This is being done through the use of N4L information objects that are uniquely and persistently identified with Digital Object Identifiers. By embedding N4L DOIs into our data structures, this will free us from the necessity of constantly updating the taxonomic information tied to each sequence; yet guarantee that the associated information is up-to-date. As the N4L model supports multiple taxonomic views and concepts, it will also provide a mechanism whereby deviations between different models exist, especially those that are constrained by rules of nomenclature. This is helping to identify areas where nomenclatural anomalies remain because rate of sequencing significantly outstrips the pace of taxonomic revision. This approach will also provide a mechanism whereby persistent links can be established to novel evolutionary lineages of critical importance to DOE missions at the earliest possible point in time, well before such lineages are subject to the formal rules of nomenclature.

References

1. Cole, J. R., B. Chai, T. L. Marsh, R. J. Farris, Q. Wang, S. A. Kulam, S. Chandra, D. M. McGarrell, T. M. Schmidt, G. M. Garrity, and J. M. Tiedje. 2003. "The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy," *Nucleic Acids Res* 31:442-3.
2. Garrity, G. M., and C. Lyons. 2003. "Future-proofing biological nomenclature," *OMICS* 7:31-31.
3. Lilburn, T. G., and G. M. Garrity. 2003. "Exploring prokaryotic taxonomy.,"*Int. J. System. Evol. Micro.* 53:7-13
4. Garrity, G. M., and T. G. Lilburn. 2005. "Self-organizing and self-correcting classifications of biological data," *Bioinformatics* 21:2309-14.

* Presenting author

7

Geobacter Project Subproject I: Genome Sequences of *Geobacteraceae* in Subsurface Environments Undergoing *In Situ* Uranium Bioremediation and on the Surface of Energy-Harvesting Electrodes

Jessica Butler* (jbutler@microbio.umass.edu), Regina O'Neil, Dawn Holmes, Muktak Aklujkar, Ray DiDonato, Shelley Haveman, and **Derek Lovley**

University of Massachusetts, Amherst, MA

The overall goal of the Genomics:GTL *Geobacter* Project is to develop genome-based *in silico* models that can predict the growth and metabolism of *Geobacteraceae* under a variety of environmental conditions. These models are required in order to optimize practical applications of *Geobacteraceae* that are relevant to DOE interests. The goal of Subproject I is to determine the genetic potential of the *Geobacteraceae* present in subsurface environments undergoing *in situ* uranium bioremediation and on the surface of energy-harvesting electrodes. This not only provides information on what metabolic modules need to be included in the *in silico* models but makes it possible to monitor the metabolic state and rates of metabolism in diverse environments by measuring transcript levels of key diagnostic genes. In order to be as comprehensive as possible, the plan is to obtain genome sequences from: 1) pure culture isolates for which there is substantial physiological information; 2) isolates from environments of interest that have the 16S rRNA gene sequences identical to those of the *Geobacter* species that predominate in these environments; 3) single cells from the environments of interest; and 4) genomic DNA extracted from the environments of interest. Progress has been made with all four approaches.

The availability of several new *Geobacteraceae* genome sequences has made it possible to further evaluate the diversity within sequences of closely related members of this family. For example, an in-depth comparison of the relatively well-studied genome of *Geobacter sulfurreducens* and the recently completed genome of *Geobacter metallireducens* revealed that only 60% (2131) of all genes were orthologous between the two genomes. The largest region of the *G. metallireducens* genome that is not present in *G. sulfurreducens* encoded genes for the degradation of aromatic compounds, consistent with the capacity for aromatics metabolism by *G. metallireducens*, but not *G. sulfurreducens*. Although both organisms contain genes for ca. 125 *c*-type cytochromes only 55% of the cytochromes in *G. sulfurreducens* have orthologs in the *G. metallireducens* genome. There are many instances of species-specific duplications of cytochrome genes, as well as instances of gain or loss of heme-binding motifs in orthologous cytochromes. Surprisingly, *c*-type cytochromes that have been shown to be important in Fe(III) reduction in *G. sulfurreducens*, such as the outer-membrane cytochromes OmcB, OmcF, OmcG, and OmcS, do not have orthologs in the *G. metallireducens* genome. Periplasmic cytochromes appear to be better conserved. For example, there are orthologs to the two periplasmic cytochromes, PpcA and MacA, previously shown to be important in Fe(III) reduction in *G. sulfurreducens*. There was much higher conservation of cytoplasmic proteins. For example, 87% of the 589 proteins in the current *in silico* model of central metabolism of *G. sulfurreducens* have orthologs in *G. metallireducens*. Of the 136 cytoplasmic genes that appear to be involved specifically in the oxidation of acetate and electron transport to the inner membrane in *G. sulfurreducens*, 92% have orthologs in *G. metallireducens*. This is significant because acetate is the key electron donor driving *in situ* uranium bioremediation and production of electricity. These findings suggest that modeling of the central metabolism of the environmentally relevant *Geobacter* species may benefit from the existing *G. sulfurreducens in silico* model. However, the results also suggest that there may be little conservation among *Geobacter*

* Presenting author

species in the proteins involved in electron transfer through the periplasm and outer membrane, with the exception of the electrically conductive pilin nanowires that are thought to be the conduit for electron flow from the surface of the cell onto Fe(III) oxides. This conclusion was also supported by analysis of the recently completed genome of *Pelobacter carbinolicus* and the draft genome of *Pelobacter propionicus.*

The first milestone towards the goal of sequencing the genomes of predominant *Geobacteraceae* in the environments of interest has been reached. A 4.8 Mbp draft genome sequence is now available for *Geobacter uraniumreducens*, which was isolated from the in situ uranium bioremediation study site in Rifle, Colorado with an environment-simulating medium in which clay-size minerals served as the source of Fe(III). The 16S rRNA gene sequence of *G. uraniumreducens* matches a sequence that predominates during in situ uranium bioremediation, and it is expected that the short interval between isolation and genome sequencing minimized any potential inactivation, loss, and rearrangement of genes that can accompany propagation of a strain in the laboratory. Although *G. uraniumreducens* has genes for 100 c-type cytochromes, only 20 multi-heme cytochromes have orthologs in the *G. sulfurreducens* and *G. metallireducens* genomes. This contrasts with the presence of a high proportion of orthologs for central metabolism and inner membrane electron transport enzymes.

This year, techniques were developed for immediately freezing samples in the field such that the cells remain intact for subsequent cultivation or sequencing of single cells. With this method it has been possible to recover additional organisms from the uranium bioremediation study site with 16S rRNA gene sequences that are identical to those that predominate during in situ uranium bioremediation. Furthermore, using a multiple displacement amplification approach, genomic DNA was amplified from single cells of *Geobacter* species obtained from these samples. We are awaiting the sequencing results.

Quantification of the transcript levels of a broad range of *Geobacteraceae* genes in a diversity of subsurface environments requires information on the sequence heterogeneity among genes that are diagnostic of important metabolic states. Although some information can be obtained from the genome sequencing described above, these approaches can not yet reasonably be applied to a large number of environments or at a large number of time intervals during the in situ uranium bioremediation process. The sequence diversity of key diagnostic genes was characterized in detail for four sedimentary environments. Degenerate PCR primers were designed from the sequences of genes that are highly conserved throughout the range of pure culture *Geobacteraceae*, as well as the genomic DNA from environments in which *Geobacteraceae* predominate. The diversity of 16S rRNA gene sequences detected was extremely low both within and among these sites. *Geobacteraceae* with 16S rRNA gene sequences closely related to the subsurface isolate *Geobacter bemidjiensis* accounted for 50-98% of the microbial community, and these sequences were 97-100% similar to each other. However, other genes amplified with this method had sequence similarities as low as 50-75%. These findings have a significant impact on strategies for evaluating gene expression in *Geobacteraceae*-dominated environments.

8

Geobacter Project Subproject III: Functional Analysis of Genes of Unknown Function

Maddalena Coppi* (mcoppi@microbio.umass.edu), Carla Risso, Gemma Reguera, Ching Leang, Helen Vrionis, Richard Glaven, Muktak Aklujkar, Xinlei Qian, Tunde Mester, and **Derek Lovley**

University of Massachusetts, Amherst, MA

The development of models that can predict the physiological responses of *Geobacteraceae* under different environmental conditions in contaminated subsurface environments or on the surface of energy-harvesting electrodes requires an understanding of the function of the genes expressed in these environments. However, no function has been assigned to a substantial number of genes in *Geobacteraceae* genomes and the actual physiological role of many genes annotated as having specific physiological functions has yet to be assessed.

For example, it is important to understand acetate metabolism in *Geobacter* species, because acetate is the electron donor driving *in situ* uranium bioremediation and electricity production. Analysis of the *Geobacter sulfurreducens* genome revealed the presence of three homologs of the monocarboxylate transporter of *E. coli*, YjcG, which has recently been demonstrated to catalyze sodium-dependent acetate uptake. These homologs are 54-56% similar to *E. coli* YjcG and are *ca.* 90% similar to each other. *G. sulfurreducens* retained the ability to grow on acetate if the transporter genes were deleted singly, but a triple mutant could not be isolated. This result suggests that the three transporters are essential for growth on acetate, but that they may be functionally redundant. We are currently investigating the possibility of compensatory interactions between the three transporters in the mutant strains. The central metabolic pathways for acetate oxidation and incorporation into biomass have also been subjected to intensive genetic analysis. A model of acetate metabolism based on the results of these studies will be presented.

Previous studies demonstrated that at subatmospheric oxygen tensions, *G. sulfurreducens* can grow utilizing oxygen as an electron acceptor, a physiological capability that may be important for survival of *Geobacter* species in subsurface environments. Two complexes potentially involved in oxygen respiration are encoded in the genome of genome of *G. sulfurreducens*, a cytochrome *c* oxidase and a cytochrome *d* oxidase. Both enzymes are present in most aerobic bacteria and have a low or high affinity for oxygen, respectively. A *G. sulfurreducens* mutant lacking the cytochrome *c* oxidase could not grow on oxygen, but retained the ability to consume oxygen, presumably via the cytochrome *d* oxidase. Further genetic analysis of this possibility is underway.

The physiologic role of the SfrAB complex of *G. sulfurreducens*, which was previously designated a cytoplasmic Fe(III) reductase, was investigated in detail, because understanding the site of metal reduction is crucial for modeling the energetics of this process. A knockout mutant deficient in SfrAB could not grow with acetate as the electron donor with either fumarate or Fe(III) as the electron acceptor, but readily grew with hydrogen or formate as the electron donor, if acetate was provided as a carbon source. This phenotype suggested that the SfrAB-deficient mutant was specifically impaired in acetate oxidation via the TCA cycle. After several weeks, SfrAB deficient strains developed the ability to grow on fumarate with acetate serving as the electron donor. Membrane and soluble fractions prepared from these acetate-adapted strains, were depleted of NADPH-dependent Fe(III), viologen, and quinone reductase activities relative to those of wild type. It was hypothesized that the lack of SfrAB inhibits growth on acetate by increasing the NADPH:NADP

* Presenting author

ratio and thereby inhibiting the isocitrate dehydrogenase reaction of the TCA cycle. Comparison of global gene expression profiles in the adapted mutant and wild type strains provided evidence of ATP depletion, impaired amino acid biosynthesis, and decreased rates of acetate uptake, all of which were consistent with a suboptimal rate of acetate oxidation via the TCA cycle in the mutant. In addition, a potential NADPH-dependent ferredoxin oxidoreductase was upregulated in the mutant. These results indicate that SfrAB is not an Fe(III) reductase, but rather might serve as a major route for NADPH oxidation in *G. sulfurreducens*. These findings, coupled with the fact that metabolically active *G. sulfurreducens* spheroplasts were incapable of Fe(III) reduction suggest that cytoplasmic Fe(III) is not an important process in *G. sulfurreducens*, consistent with other recent functional studies.

Functional analyses have also provided new insights into the mechanisms by which *G. sulfurreducens* produces electricity. We have recently discovered that *G. sulfurreducens* expresses electrically conductive pili that appear to serve as electrical conduits from the cell to Fe(III) and Mn(IV) oxides. As reported last year, initial genetic studies suggested that that outer-membrane *c*-type cytochrome, OmcS, mediated electrical contact between *G. sulfurreducens* and that pili were not required for electricity production. These results may have been due to the fact that, in these studies, power production was limited by the rate of electron transfer from the cathode to oxygen in the overlying water. When cathode limitation was eliminated via the use of a potentiostat, current production by wild type *G. sulfurreducens* increased more than 10-fold and the pilus-deficient mutant was found to be significantly impaired, with a power output that was only 17% of that of wild type. Confocal laser scanning microscopy revealed that, in the absence of cathode limitation, the wild-type cells produced thicker biofilms on the anode, whereas the pilus-deficient mutant produced thinner biofilms more similar to those observed in the previously used cathode-limited systems. These results suggest that cells that are not in close contact with the electrode may transfer electrons through the biofilm via the electrically conductive pili. Deletion of a gene designated *gumC* eliminated the production of exopolysaccharide and also negatively impacted power production, reducing it to less then half of wild type levels. These results suggest that exopolysaccharide production also plays an important role in electron transfer to electrodes. A combination of biochemical, genetic and microscopic analyses demostrated that the *gumC* mutant overproduces pili, which may enable it to compensate for the absence of exopolysaccharide production and transfer electrons to insoluble electron acceptors. Production of current by a *gumCpilA* double mutant was less than 10% of wild type.

Numerous other functional genomics studies are underway. For example, genetic analysis of putative metal resistance genes lead to the identification of genes involved zinc, cadmium, and copper resistance. Surprisingly, deletion of one of the copper-resistance genes, *cusA*, also prevented growth with Fe(III) serving as the electron acceptor. Additional outer membrane *c*-type cytochromes have been implicated in electron transfer to Fe(III) citrate, Fe(III) oxides and electrodes by genetic studies, but several of these cytochromes appear to have functions that are not directly related to electron transfer to metals or electrodes. Additional targets for functional analysis have been selected based on higher levels of expression during growth on Fe(III) oxide or electrodes. The function of proteins that form complexes with cytochromes previously shown to be involved in extracellular electron transfer are also being investigated as are those of proteins which are conserved throughout the *Geobacteraceae*, but not found in the genomes of other organisms.

9

Comparative Genomic Analysis of Five *Rhodopseudomonas palustris* Strains: Insights into Genetic and Functional Diversity within a Metabolically Versatile Species

Yasuhiro Oda[1]* (yasuhiro@u.washington.edu), Frank W. Larimer[2], Patrick Chain[3], Stephanie Malfatti[3], Maria V. Shin[3], Lisa M. Vergez[3], Loren Hauser[2], Miriam L. Land[2], Dale A. Pelletier[2], and **Caroline S. Harwood**[1]

[1]University of Washington, Seattle, WA; [2]Oak Ridge National Laboratory, Oak Ridge, TN; and [3]Lawrence Livermore National Laboratory, Livermore, CA

Rhodopseudomonas palustris is a facultatively photosynthetic bacterial species that has the potential to be used as a biocatalyst for hydrogen production, carbon sequestration, biomass turnover, and biopolymer synthesis. The genome of *R. palustris* strain CGA009 has been reported and consists of a 5.46 Mb chromosome with 4836 predicted protein-coding genes. Several studies have shown that the *R. palustris* species is comprised of genetically and phenotypically diverse strains. To identify the core characteristics of the species that are essential for proper physiological functioning and to identify new metabolic capabilities, the DOE Joint Genome Institute sequenced four additional strains of *R. palustris.* These strains, BisB5, HaA2, BisB18, and BisA53, were directly isolated from agar plates that had been inoculated with freshwater sediments. Their 16S rRNA gene sequences differ from that of strain CGA009 by about 2% and their BOX-PCR genomic DNA fingerprint patterns differ significantly. The genome of strain BisB5 consists of a 4.89 Mb chromosome with 4386 predicted genes. Strain HaA2 has a 5.33 Mb chromosome with 4687 predicted genes, strain BisB18 has a 5.51 Mb chromosome with 4949 predicted genes, and strain BisA53 has a 5.50 Mb chromosome with 4913 predicted genes. Approximately 60 to 80% (depending on the strain) of the genes from strain CGA009 were present in each strain, and these may represent the core genes of the *R. palustris* species. However, whole genome comparisons among strains showed a high degree of genome rearrangement in terms of gene orders and reading directions. Furthermore, there were high numbers of genes (250 to 560 genes) that were specific to a given strain and not seen in any other strain. Based on their gene inventories, each strain is predicted to have strain-specific physiological traits. Strain CGA009 is well equipped for nitrogen fixation with three nitrogenase isozymes and four sets of glutamine synthetases, strain BisB5 has expanded anaerobic aromatic degradation capabilities (e.g., phenylacetate degradation), strain HaA2 should be well-adapted for growth in oxygen as it encodes seven different aerobic terminal oxidases, and strain BisB18 should be able to grow well anaerobically in dark (e.g., carbon-monoxide dehydrogenase genes, three sets of pyruvate-formate lyase genes, formate-hydrogen lyase genes, and DMSO reductase genes). Finally, strain BisB53 has an expanded set of exopolysaccharide synthesis genes and readily attaches to surfaces to form biofilms. Despite these differences, there were relatively few obvious examples of lateral gene transfer in the genomes and the genomes harbor relatively few insertion sequences or transposons compared to other bacterial species. Our comparative genomic analysis suggests that *R. palustris* is a dynamic species comprised of diverse ecotypes that are well adapted to specific environmental niches.

* Presenting author

10

Functional Analysis of *Shewanella*, a Cross Genome Comparison

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

Marine Biological Laboratory, Woods Hole, MA

Shewanella oneidensis MR-1 was initially chosen as a model organism by the Department of Energy (DOE) based on its unique metal reducing and bioremediation capabilities. Data generated from the analyses of various *Shewanella* strains showed that this genus contained species adapted to life in a variety of environmental niches including land, lake sediments, fresh water and marine environments. The range of environments where *Shewanella* successfully lives implies that it is highly versatile in its capacity to respire, metabolize nutrients, and sense its surroundings for available nutrient sources and electron acceptors. DOE subsequently funded the sequencing of several additional *Shewanella* strains of varying ecological properties, and high quality draft sequences of these genomes have been made available. In our work we analyze the predicted protein sequences from *S. oneidensis* MR-1, 10 strains sequenced at JGI (*S. putrefaciens* CN-32, *S. loihica* (formerly alga) PV-4, *S. baltica* OS155, *S. frigidimarina* NCIMB400, *S. denitrificans* OS217, *S. amazonensis* SB2B, *Shewanella* sp. MR-7, *Shewanella* sp. ANA-3, *Shewanella* sp. MR-4, *Shewanella* sp. W3-18-1), and two environmental samples from the Sargasso Sea (*Shewanella* SAR-1, *Shewanella* SAR2) sequenced by TIGR.

Our research focuses on studying groups of sequence related proteins and whether such groups can give us insight into how organisms adapt to their environments. The duplication of genes followed by diversification of their sequences results in a group of proteins encoding similar or related functions. This process is believed to be an important means of functional specialization and adaptation. The *Shewanella* genome sequences provide an excellent resource to study the distribution of protein groups and to analyze the activities they encode in order to find evidence of specialization of functions related to their metabolic capabilities and their environmental phenotypes.

Pair-wise alignments of the protein sequences encoded by the *Shewanella* genomes were produced using the AllAllDb program of Darwin (Data Analysis and Retrieval With Indexed Nucleotide/peptide sequence package), version 2.0. Fused proteins (arising from gene fusion events) were separated into smaller proteins corresponding to their un-fused components. Protein groups were then generated from the pair-wise alignments in a transitive grouping process. Two methods were used to compare protein groups across the 13 *Shewanella* genomes. In one method groups were initially generated from the proteins encoded by *S. oneidensis* MR-1., and these groups were further used to search for sequence similar matches in the other 12 genomes. Sequence similarities of 175 PAM units or less and alignments over at least 45% of the protein sequences were applied for this comparison. A total of 406 protein groups containing 2 or more *S. oneidensis* MR-1 proteins were compared this way. In the second method sequence related groups were generated directly from the pair-wise alignments of the proteins from the 13 *Shewanella* strains. The same transitive grouping process was applied, but the sequence similarities were restricted to 125 PAM units or less over 70% of the protein sequences. In the second method 4702 protein groups were generated.

We are analyzing the protein groups for their distribution among the 13 *Shewanella* strains and for the functions they encode. Groups of proteins with functions relating to anaerobic respiration, chemotaxis and environmental sensing will be presented.

* Presenting author

11

Comparative Genomic and Proteomic Insight into the Evolution and Ecophysiological Speciation in the *Shewanella* Genus

James M. Tiedje*[1] (tiedjej@msu.edu), Konstantinos T. Kostantinidis[1], Joel A. Klappenbach[1], Jorge L.M. Rodrigues[1], Mary S. Lipton[4], Margaret F. Romine[4], Sean Conlan[10], LeeAnn McCue[4], Patrick Chain[6], Anna Obraztsova[3], Loren Hauser[2], Margrethe Serres[7], Monica Riley[7], Carol S. Giometti[5], Eugene Kolker[8], Jizhong Zhou[2], Kenneth H. Nealson[3], and James K. Fredrickson[4]

[1]Michigan State University, East Lansing, MI; [2]Oak Ridge National Laboratory, Oak Ridge, TN; [3]University of Southern California, Los Angeles, CA; [4]Pacific Northwest National Laboratory, Richland, WA; [5]Argonne National Laboratory, Argonne, IL; [6]Lawrence Livermore National Laboratory, Livermore, CA; [7]Marine Biological Laboratory, Woods Hole, MA; [8]BIATECH Institute, Bothell, WA; [9]University of Oklahoma, Norman, OK; and [10]Wadsworth Center, Troy, NY

Members of the genus *Shewanella* are found in a variety of environments, such as freshwater lakes, marine sediments, subsurface formations, and at variable depths in redox stratified aquatic systems. The ecological and physiological diversities among species of this genus suggest a high degree of specialization. The mechanisms and factors that drive speciation are still not well understood. For this reason, we are taking advantage of full genome sequencing of 17 *Shewanella* species and aim to: 1) define the genetic differences and mechanisms that account for the ecological success in different environments and support different physiologies, 2) identify mechanisms of evolution and speciation for *Shewanella* species, and 3) determine a set of core genes important for metal reduction. Results from four genomic sequences (*S. frigidimarina*, *S. putrefaciens* CN-32, S. sp. PV-4, and *S. denitrificans*) compared to the finished genome of *S. oneidensis* MR-1 indicated the presence of 10,000 unique genes, while only 2,200 genes are shared among all sequenced species. Each sequenced species contained a subset of genome specific genes (25%), with a substantial number of those (16%) annotated as hypothetical open reading frames (ORF). The pair-wise genetic distances using the average nucleotide identity (ANI) of all genes according to the method of Kostantinidis and Tiedje (2005) between these 4 genomes were approximately 70%, indicating that these genomes are not closely related, and hence a species might harbor several distinct ecotypes, appropriately evolved for a specific environment. The above 4 genomes together with an additional 7 that have been more recently sequenced to offer resolution within species present an unprecedented evolutionary gradient for study of the diversification of a bacterial genus at varied level of resolution. The sequenced genomes with various degree of relatedness within a single genus will provide ideal model system for studying evolutionary processes and forces such as positive, neutral and negative selection in prokaryotes (Figure 1).

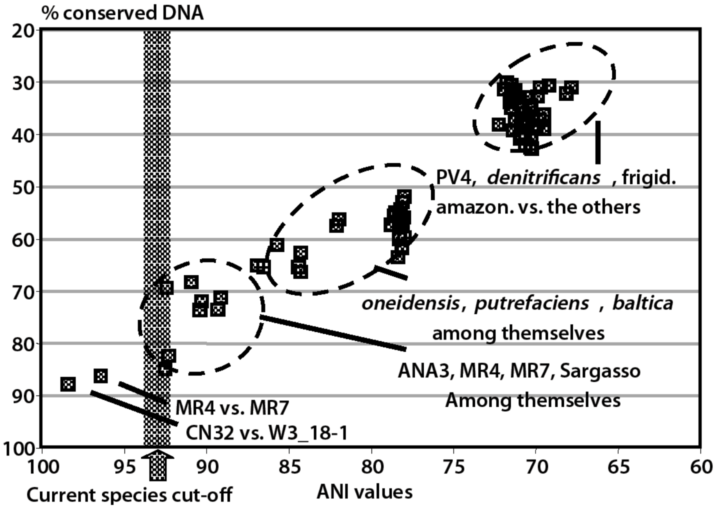


Fig. 1. Conserved gene core vs. evolutionary distance.

* Presenting author

Insights from the analysis of 11 *Shewanella* genomes indicated that 2,000-2,200 genes are shared with *Vibrio*, a number higher than those shared within the *Shewanella* genus. This finding demonstrates the challenges of identifying ecologically relevant genes, based solely on sequence analysis. Thus, we are analyzing the proteomes of all different species to identify a core set of expressed proteins under defined conditions. Whole cell lysates from 13 *Shewanella* species were analyzed by two-dimensional electrophoresis (2DE) and the patterns were compared. Sorting the 2DE patterns by constellations of similar spots resulted in grouping of the species in close approximation to the *gyrB*-based phylogenetic tree. We are studying the distribution of sequence similar protein families in the 13 *Shewanella* genomes to detect functional adaptation relating to their environments. We are also examining the species distribution of cytochromes, short repeats, and IS elements, as well as conservation of regulatory elements for insights into the evolution of these species. In addition, physiological studies for different strains are underway, testing salinity, temperature, pH, carbon sources under aerobic conditions, a variety of electron acceptors with lactate or N-acetyl glucosamine as electron donor, ability of producing current in a mediator-less biofuel cell, and growth on Cr(VI). Physiological studies will be correlated to the genomic and proteomic content and variability in all species.

Reference

1. Kostantinidis, K. and J.M. Tiedje. 2005. "Genomic insights that advance the species definition for prokaryotes." *Proc. Natl. Acad. Sci. U.S.A.* 7:2567-2572.

12

Microbial Genome Sequencing with Ploning from the Wild: A Progress Report

Kun Zhang[1]* (kzhang@genetics.med.harvard.edu), Adam C. Martiny[2], Nikkos B. Reppas[1], Kerrie W. Barry[3], Joel Malek[4], Sallie W. Chisholm[2], and **George M. Church**[1]

[1]Harvard Medical School, Boston, MA; [2]Massachusetts Institute of Technology, Cambridge, MA; [3]Joint Genome Institute, Walnut Creek, CA; and [4]Agencourt Bioscience, Beverley, MA

With less than 1% of microorganisms easily cultured, obtaining genome sequence and continuity for the remaining 99% has been one of the greatest challenges in environmental genomic studies. We aimed at developing a method, polymerase cloning (ploning) for genome sequencing directly from single uncultured cells, and applying the method to study the Prochlorococcus community structure in open oceans.

The first critical component of ploning is to perform whole genome amplification on a single template molecule. We have reported a real-time isothermal amplification system that successfully addresses the issue of background non-specific amplification in last year's meeting. We also investigated amplification bias in two *E. coli* polymerase clones (plones) using Affymetrix chip hybridization, and showed that bias is randomly distributed.

Here we present several recent progresses in the second critical component of our method: genome sequencing from plones. We have prepared several plones from single Prochlorococcus cells of the MIT 9312 strain. Our initial attempt of performing shotgun sequencing on such plones failed because of issues in sequencing library construction, such as abnormal insert size, high vector content and low cloning efficiency. We hypothesized that such problems are due to the high order DNA

* Presenting author

branching structure generated by multiple displacement amplification. A three-step enzymatic treatment has been developed to resolve the high order DNA structure, so that the chimeric rate has been reduced from as high as 52% to 6%. We have sequenced two MIT 9312 plones, one at the sequencing depth of 3.5x and the other at 4.7x, and recovered 62% and 66% of the genome respectively. Full genome coverage can be achieved by increasing the sequencing depth to ~24x or PCR sequencing from the plones. The estimated mutation rate in single cell amplification is $<2\times10^{-5}$.

When applying the ploning method to ocean samples collected from Hawaii, we encounter another technical problem: the amount of cell-free DNA is more than that within a live cell in single-cell dilution. We have developed a DNase-based protocol to remove contamination of cell-free DNA. We are screening for good *Prochlorococcus* plones from ocean samples, and will sequence a few in the near future.

13

Application of a Novel Genomics Technology Platform

Karsten Zengler[1], Marion Walcher[1], Carl Abulencia[1], Denise Wyborski[1], Trevin Holland[1], Fred Brockman[2], Cheryl Kuske[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

A technology platform has been developed to obtain whole genome sequences from targeted uncultured microorganisms. Our approach combines fluorescence in situ hybridization (FISH) followed by amplification of whole genomes using multiple displacement amplification (MDA). Microcolonies in microcapsules derived from high throughput cultivation (HTC) or individual cells from the environment are specifically stained with fluorescently labeled oligonucleotide probes targeting 16S rRNA. Target cells are further isolated from non-target microorganisms by flow cytometry. Genomes of these isolated cells are subsequently amplified by whole genome amplification.

The genome from one *Acidobacteria*-microcolony was amplified by MDA. Furthermore, positive MDA products were obtained from 5, 50 and 100 *Acidobacteria* cells and 100 cells affiliated to candidate division TM7 isolated directly from the environment. A small insert library from a MDA product derived from 50 *Acidobacteria* cells has been constructed and a portion of the library (~1000 clones) has been sequenced. The library appears to predominantly contain sequences from an *Acidobacteria* division member that represents a major uncultured subgroup (group #6). Blast scores to published sequences were generally very low, but several clones were directly affiliated to *Solibacter usitatus*, a group 3 Acidobacterium. A total of 82 contigs were assembled which were derived from 982 sequences. The largest of these contigs consists of over 6 kb and was derived from 27 sequences. Three of those sequences are closely related to *Solibacter usitatus*.

* Presenting author

Section 2
Microbial-Community Sequencing

14
Ribosomal Database Project II: Sequences and Tools for High-Throughput rRNA Analysis

J.R. Cole* (colej@msu.edu), B. Chai, R.J. Farris, Q. Wang, S.A. Kulam, D.M. McGarrell, G.M. Garrity, and **J.M. Tiedje**

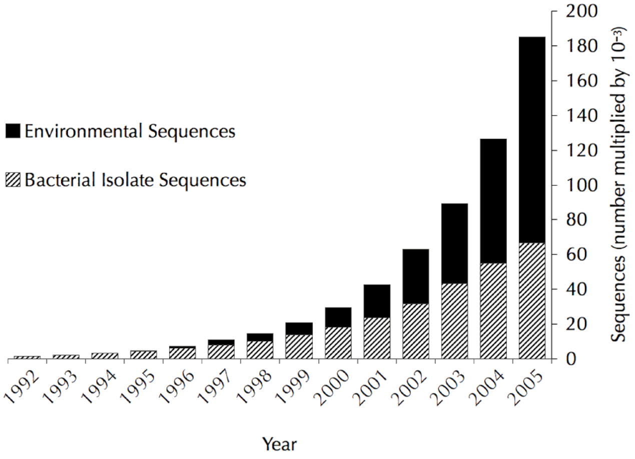
Michigan State University, East Lansing, MI

The Ribosomal Database Project II (RDP) provides data, tools, and services related to ribosomal RNA sequences to the research community (Cole et al., 2005. Nucleic Acids Res 33:D294). Through its website (http://rdp.cme.msu.edu), RDP-II offers aligned and annotated rRNA sequence data and analysis services. These data and services help discovery and characterization of microbes important to energy, biogeochemical cycles and bioremediation.

The RDP-II databases are updated monthly with data obtained from the International Nucleotide Sequence Databases (GenBank/EMBL/DDBJ). As of October 2005 (Release 9.31), RDP-II maintains 184,990 aligned and annotated bacterial small-subunit rRNA gene sequences (Figure 1). These sequences are available in several subsets, including higher quality near-full-length sequences (72,540), sequences from environmental samples (118,450), from in-culture bacterial isolates (66,540), and sequences from bacterial species type strains (4,445). The latter are of special interest, because species type-strains serve as archetype and link rRNA-base phylogeny with bacterial taxonomy. These type-strain sequences cover about 60% of the validly named bacterial species.

High-throughput environmental rRNA projects routinely produce hundreds to thousands of sequences per sample. The RDP-II tools have been redesigned to accommodate this trend towards large-scale rRNA sequencing efforts. Among the tools offered in RDP Release 9, Hierarchy Browser allows users to rapidly navigate through the RDP sequence data, RDP Classifier provides a rapid taxonomic placement of one to hundreds of user sequences, Sequence Match (completely re-implemented for Release 9) is more accurate than BLAST at rapidly finding closely related rRNA sequences, and the new version of Probe Match finds probe and primer binding sites using a more efficient algorithm and enables users to skip partial

Figure 1. Increase in number of publicly available bacterial small-subunit rRNA sequences.



* Presenting author

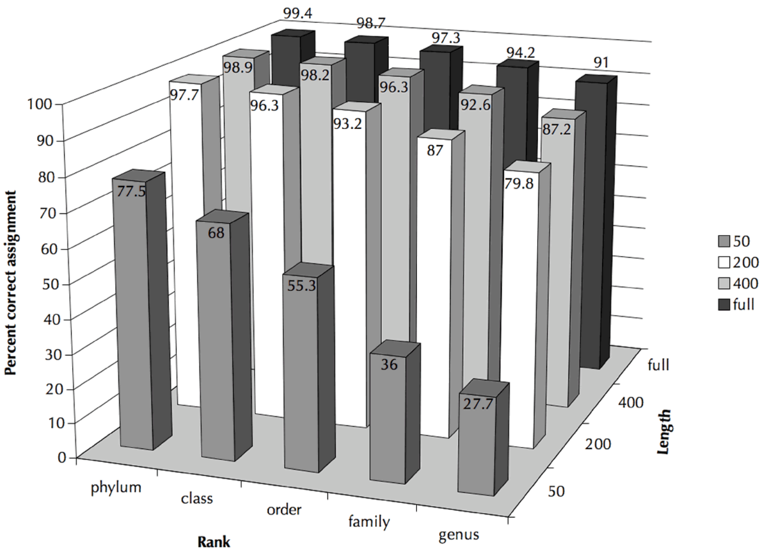
sequences missing the target region. The newly developed tool, Library Compare, combines the RDP Classifier with a statistical test to flag taxa differing significantly between 16S rRNA gene libraries.

The RDP Classifier was trained using information for approximately 5,486 bacterial species type strains (and a small number of other sequences representing regions of bacterial diversity with few named organisms) in 896 genera. One recent study found up to 5% of rRNA sequences examined contained sequencing artifacts when examined using a novel method (Ashelford et al., 2005. Appl Environ Microbiol 71:7724). This study included most of our training set sequences but found potential artifacts in only three, underscoring the importance of using well-characterized sequences for rRNA comparisons.

For the near full-length and 400 base partial rRNA sequences, the RDP Classifier was accurate down to the genus level (Figure 2), while even with 200 base partial sequences the RDP Classifier was accurate down to the family levels. The RDP Classifier did not perform well for partial sequences of length 50, likely due to insufficient features provided by such short partial sequences.

Over 60% of the full-length sequences misclassified at the genus level were more similar to one or more sequences derived from members of other genera than to other sequences within the same genus. Although some of these may be due to mistakes in sequence provenance, it seems likely that many of these "misclassifications" represent divergence in the nomenclatural taxonomy from the underlying (rRNA based) phylogeny. We are collaborating with the Taxomatic Project (Lilburn & Garrity 2004. Int J Syst Evol Microbiol 54:7) to help correct these discrepancies.

Figure 2. RDP Classifier accuracy by query size. (Exhaustive leave-one-out testing.)



* Presenting author

15

Community Genomics as the Foundation for Functional Analyses of Natural Microbial Consortia

R.J. Ram[1], V.J. Denef[1], I. Lo[1], N.C. VerBerkmoes[2], G. Tyson[1], G. DiBartolo[1], E.A. Allen[1], J. Eppley[1], B.J. Baker[1], M. Shah[2], R.L. Hettich[2], M.P. Thelen[3], and **J.F. Banfield**[1]* (jill@seismo.berkeley.edu)

[1]University of California, Berkeley, CA; [2]Oak Ridge National Laboratory, Oak Ridge, TN; and [3]Lawrence Livermore National Laboratory, Livermore, CA

The objective central to our DOE Genomics:GTL project is develop methods for the study of microbial communities in their natural environments. In order to circumvent barriers associated with cultivation and limitations that arise from studies of isolated organisms, we are developing methods for genomic and functional characterization of microbial communities *in situ*. Genomic data provide the information needed to monitor activity as organisms assemble to form consortia and respond to perturbations in their environments. As our approach relies upon the availability of relatively comprehensive genomic information for the dominant community members, our efforts focus on chemoautotrophic biofilms with low species richness that grow in the subsurface in association with a dissolving metal sulfide deposit. The self-sustaining microbial communities that populate acid mine drainage systems are particularly amenable to high-resolution ecological analyses. In addition to their utility as model systems, these communities are important targets for study because microbially promoted dissolution of metal sulfides causes acid mine drainage, a major environmental problem associated with energy resources. It is also a process that underpins bioleaching-based metal recovery and coal desulfurization and mercury removal. Samples are collected from a range of locations within the Richmond mine at Iron Mountain, Redding, CA.

To date, extensive community genomic data have been obtained from an air-solution interface biofilm and a subaerial biofilm. Although both biofilms are dominated by *Leptospirillum* group II species that appear clonal at most loci, detailed analyses reveal significant heterogeneity in gene content in certain genomic hot spots. Variability in gene content occurs both within and between populations. Furthermore, some large genomic blocks exhibit anomalously high sequence identity, suggesting very recent lateral transfer of genomic regions between the populations. Similar observations have been made for the archaeal populations.

In the first study of its type, genomic data from one biofilm were used to characterize the proteome of a similar biofilm using mass spectrometry-based proteomics. Over 2,000 proteins were confidently identified, allowing initial insights into partitioning of function and the environmental challenges faced by the microbial community (Ram et al. 2005). As it is impractical to acquire extensive genomic data from every site sampled, we have used this and subsequent datasets to evaluate the influence of genome sequence dissimilarity on the efficiency of peptide and protein detection. Specifically, we are testing the likelihood of cross detection of proteins from closely related microbial species and evaluating the limitations imposed by sequence divergence for detection of proteins from closely related strains.

We developed an end member random substitution model that considers the average amino acid dissimilarity, the average peptide length, the fraction of uniquely detectable peptides, and predicted protein length. The model predicts a preferential loss of shorter proteins and a sigmoid relationship between the fraction of proteins detected and the average amino acid dissimilarity. Because amino acid substitutions do not occur randomly, our analyses also make use of genomic data from closely

* Presenting author

related strains and species. To date, results suggest high protein detection rates using genomic data for proteome analysis of closely related strains, significant discrimination of proteins from different species, and highlight the importance of identifying unique peptides for distinguishing proteins from strain variants and closely related species.

For abundant proteins with regions of no peptide coverage, we PCR amplified the corresponding gene directly from the environmental sample and determined that there were very few cases where amino acid changes caused a peptide to be undetected, even in cases where abundant proteins had very divergent coding regions at the nucleotide level (up to 7%). This result is generally consistent with predictions from modeling.

For the environmental *Leptospirillum* group II populations and a new isolate, gene variants often have >20 nucleotide changes but few amino acid substitutions. Environmental samples tend to be dominated by populations with a single sequence type for most genes. Variants at each locus segregated independently among different populations, suggesting that there may be a few ancestral strains that have recombined to produce the extant populations. Although the bacterial groups appear to be shaped by recombination, DNA transfer among closely related organisms appears less common than in coexisting archaea, where we have determined that homologous recombination represents a major force in population dynamics.

Research sponsored by the Genomics:GTL program, Office of Biological and Environmental Research, U.S. Department of Energy.

Reference

1. Ram, R.J., VerBerkmoes, N.C., Thelen, M. P., Tyson, G.W., Baker, B.J. Blake, R.C. II, Shah, M., Hettich, R.L. and Banfield, J.F. (2005) "Community proteomics of a natural microbial biofilm," *Science*, 308, 1915-1920.

16

Environmental Whole-Genome Amplification to Access Microbial Diversity in Contaminated Sediments

Denise L. Wyborski[1,3], Carl B. Abulencia[1,3], Joseph A. Garcia[1], Mircea Podar[1], Wenqiong Chen[1,3], Sherman H. Chang[1], Hwai W. Chang[1], Terry C. Hazen[2,3], and Martin Keller[1,3]* (mkeller@diversa.com)

[1]Diversa Corporation, San Diego, CA; [2]Lawrence Berkeley National Laboratory, Berkeley, CA; and [3]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

Low biomass samples from nitrate and heavy metal contaminated soils yield DNA amounts which have limited use for direct, native analysis and screening. Multiple displacement amplification (MDA) using ϕ 29 DNA polymerase was used to amplify whole genomes from environmental, contaminated, subsurface sediments. By first amplifying the gDNA, biodiversity analysis and genomic DNA library construction of microbes found in contaminated soils were made possible. We extracted DNA from samples with extremely low cell densities from a Department of Energy contaminated site. After amplification, SSU rRNA analysis revealed relatively even distribution of species across several major phyla. Clone libraries were constructed from the amplified gDNA and a small subset of clones was used for shot gun sequencing. BLAST analysis of the library clone sequences, and COG analysis, showed that the libraries were diverse and the majority of sequences had sequence identity to known proteins. The libraries were screened by DNA hybridization and sequence analysis for

* Presenting author

native histidine kinase genes. 37 clones were discovered that contained partial histidine kinase genes, and also partial, associated response regulators and flanking genes. Whole genome amplification of metagenomic DNA from very minute microbial sources enables access to genomic information that was not previously accessible.

17

Domestication of Uncultivated Microorganisms from Soil Samples

Annette Bollmann[1]* (a.bollmann@neu.edu), Lisa Ann Fagan[2], Anthony Palumbo[2], **Kim Lewis**[1] and Slava Epstein[1]

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

The majority of microorganisms from natural environments cannot be grown in the lab by standard cultivation techniques. The diffusion chamber-based approach (Kaeberlein et al. 2002, Science 296:1127) is an alternative method to grow microorganisms in a simulated natural environment. We observed (Nichols et al., unpubl.) that continuous cultivation in diffusion chambers of species initially incapable of growing in Petri dish eventually leads to production of their cultivable variants, and thus to their domestication. Here we use this approach to cultivate and domesticate microorganisms from subsurface soil samples from the FRC site (borehole FW111, area3). The microorganisms were extracted from the soil samples and incubated in diffusion chambers. The incubations took place in containers filled with wetted soil. After several weeks of incubation, the content of diffusion chambers was inoculated into a new generation of diffusion chambers, to the total of three such generations. In parallel, the chamber-grown material was inoculated into Petri dishes to monitor the process of domestication and subsequent isolation of growing microorganisms.

Phylogenetic analysis based on 16SrRNA sequences revealed clear differences between species diversity in Petri dishes inoculated with soil material as compared to Petri dishes inoculated with material from the diffusion chambers. The isolates obtained by standard approaches belong to *Alpha-* and *Gamma-Proteobacteria, Actinobacteria* and *Verrucomicrobia*. Petri dishes inoculated with diffusion chambers-grown material contained bacteria from the phyla, *Alpha-, Beta-,* and *Gamma-Proteobacteria, Actinobacteria, Firmicutes,* and CFBs. Interestingly, several isolates could only be obtained from material passaged through multiple diffusion chambers, and were never observed in the earlier generations of the chambers or in Petri dishes inoculated directly by environmental samples. Several of the isolates are close related to bacteria found in different experiments with molecular methods (North et al. 2004, AEM 70:4911; Reardon et al. 2004, AEM70:6037; Fields et al. 2005, FEMS Microb. Ecol 53:417).

In conclusion passaging environmental microorganisms through the diffusion chamber leads to the domestication of many previously uncultivated microorganisms, which enables their isolation in pure culture. The diffusion chamber-based domestication does select for specific microorganisms, the associated biases are different from those of traditional culture techniques. This provides access to cultures of at least some microorganisms previously known only by their molecular signatures.

* Presenting author

18

Metagenomic Analysis of Microbial Communities in Uranium-Contaminated Groundwaters

Jizhong Zhou[1,6,7]* (jzhou@ou.edu), Terry Gentry[1], Chris Hemme[1,6,7], Liyou Wu[1], Matthew W. Fields[2,7], Chris Detter[3], Kerrie Barry[3], David Watson[1], Christopher W. Schadt[1], Paul Richardson[3], James Bristow[3], Terry C. Hazen[4,7], James Tiedje[5], and Eddy Rubin[3]

[1]Oak Ridge National Laboratory, Oak Ridge, TN; [2]Miami University, Oxford, OH; [3]DOE Joint Genome Institute, Walnut Creek, CA; [4]Lawrence Berkeley National Laboratory, Berkeley, CA; [5]Michigan State University, East Lansing, MI; [6]University of Oklahoma, Norman, OK; and [7]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

Due to the uncultivated status of the majority of microorganisms in nature, little is known about their genetic properties, biochemical functions, and metabolic characteristics. Although sequence determination of the microbial community 'genome' is now possible with high throughput sequencing technology, the complexity and magnitude of most microbial communities make meaningful data acquisition and interpretation difficult. Therefore, we are sequencing groundwater microbial communities with manageable diversity and complexity (~10-400 phylotypes) at the U.S. Department of Energy's Natural and Accelerated Bioremediation Research (NABIR)-Field Research Center (FRC), Oak Ridge, TN. The microbial community has been sequenced from a groundwater sample contaminated with very high levels of nitrate, uranium and other heavy metals and pH ~3.7. Sequence analysis of this groundwater sample based on a 16S rDNA library revealed 10 operational taxonomic units (OTUs) at the 99.6% cutoff with >90% of the OTUs represented by an unidentified γ-proteobacterial species similar to *Frateuria*. Additional OTUs were related to a β-proteobacterial species of the genus *Azoarcus*. Three clone libraries with different DNA fragment sizes (3, 8 and 40 kb) were constructed, and 50-60 Mb raw sequences were obtained using a shotgun sequencing approach. The raw sequences were assembled into 2770 contigs totaling ~6 Mb which were further assembled into 224 scaffolds (1.8 kb-2.4 Mb). Preliminary binning of the scaffolds suggests 4 primary groupings (2 *Frateuria*-like γ-proteobacteria, 1 *Burkholderia*-like β-proteobacteria and 1 *Herbaspirillum*-like β-proteobacteria). Genes identified from the sequences were consistent with the geochemistry of the site, including multiple nitrate reductase and metal resistance genes. Despite the low species diversity of the samples, evidence of strain diversity within the identified species was observed. Analysis with functional gene arrays containing ~23,000 probes designed based on these community sequences as well as genes important for biogeochemical cycling of C, N, and S, along with metal resistance and contaminant degradation suggested that the dominant species could be biostimulated during *in situ* uranium reduction experiments at the FRC. These results also suggest that the dominant species could play a direct or indirect role in the bioremediation of uranium.

* Presenting author

19

Understanding Phage-Host Interactions Using Synergistic Metagenomic Approaches

Shannon J. Williamson* (shannon.williamson@venterinstitute.org), Douglas B. Rusch, Shibu Yooseph, Aaron Halpern, Karla Heidelberg, Cynthia Pfannkoch, Karin Remington, Robert Friedman, Marvin Frazier, Robert Strausberg, and J. Craig Venter

J. Craig Venter Institute, Rockville, MD

Marine bacteriophages (viruses that infect bacteria) are the most abundant biological entities on our planet. Interactions between phages and their hosts impact several important biological processes in the world's oceans from horizontal gene transfer to the cycling of essential nutrients. Interrogation of microbial metagenomic data collected as part of the Sorcerer II Expedition (http://www.sorcerer2expedition.org) has revealed an unexpectedly high abundance of phage sequences. Analysis of these data resulted in observations that cast new light on the nature of environmental phage-host interactions. For example, we found that the site-site distribution of phage closely related to the cyanomyophage P-SSM4 is almost identical to that of the dominant population of *Prochlorococcus* present in our samples. Studies have shown that P-SSM4 can infect divergent strains of *Prochlorococcus* (Sullivan et al. 2005), yet the co-occurrence of the P-SSM4-like phage and its host over a wide geographic area suggests a steady-state infection of the dominant *Prochlorococcus* population. It is interesting to note that the cyanomyophage P-SSP7, which is highly specific for one high light-adapted strain of *Prochlorococcus*, is largely absent from our data, suggesting the scarcity of its host at the times of sampling. As phage infection can greatly influence the clonal composition of host cell communities, our observation suggests that this particular phage may control the abundance and distribution of perhaps one of the most dominant components of picophytoplankton in oligotrophic oceans. Furthermore, environmental factors appear to influence the occurrence of temperate versus lytic phage in a hypersaline environment. Tailed phage that are members of the family *Siphoviridae* are often characterized as temperate and therefore have the ability to establish silent infections, otherwise known as lysogeny, with their hosts. By far, the greatest proportion of siphophage sequences in our dataset originated from a terrestrial hypersaline pond on Floreana Island in the Galapagos. Assembly of these sequences from the hypersaline pond resulted in the recovery of a complete phage genome approximately 50kb in length. The presence of an integrase gene confirms that this phage is indeed temperate and its genomic architecture appears to be conserved with respect to other temperate siphophage genomes. The predominance of this phage genome to the exclusion of others implies that propagation of phage through the lysogenic pathway of infection is favored over lytic replication in response to productive challenges stemming from environmental pressures. Finally, cluster analysis of viral peptide sequences revealed the presence of seven viral clusters, each containing hundreds of host-derived proteins of varying metabolic function, including certain proteins involved in the processes of photosynthesis and photoadaptation, phosphate stress and acquisition, and carbon metabolism. Distribution of the proteins within these seven clusters is geographically diverse, suggesting that viral acquisition of host metabolic genes is a more abundant and widespread phenomenon than previously recognized. In summary, our study indicates that metagenomic analysis of coincident viral and microbial sequence data provides a unique opportunity to explore phage-host interactions on a global scale.

* Presenting author

20

The Impact of Horizontal Gene Transfer: Uniting or Dividing Microbial Diversity?

Rachel J. Whitaker[1,2] (rwhitaker@nature.berkeley.edu), Dennis Grogan[3], Mark Young[4], and Frank Robb[5]

[1]University of California, Berkeley, CA; [2]University of Illinois, Urbana, IL; [3]University of Cincinnati, Cincinnati, OH; [4]Montana State University, Bozeman, MT; and [5]University of Maryland, College Park, MD

Comparative genomics has provided evidence for horizontal gene transfer (HGT) events occurring from the twigs all the way down the trunk of the tree of life. The adaptive significance of horizontal transfer events to the trajectory of microbial evolution depends upon both the frequency at which genes are exchanged and the action of natural selection upon transferred genes. For example, if horizontal transfer of homologous genes among closely related individuals within a population occurs more frequently than natural selection, gene transfer provides a cohesive force within a lineage, increasing the level of neutral diversity within a population by allowing natural selection to act at the level of the gene rather than the genome. If, on the other hand, horizontal transfer of homologous sequence is a rare, periodic selection events will purge neutral diversity and drive diversification of independent clonal 'ecotypes'. Multilocus sequence analysis and population genomics reveal that the frequent transfer of homologous sequences has a major impact on speciation and adaptive evolution in several Bacterial and Archaeal species. Recently, community and comparative genomics have revealed large differences in gene content between very closely related individuals, suggesting that the rapid movement of non-homologous sequences in and out of microbial chromosomes also drives microbial evolution. Here again, the rate of movement of novel non-homologous sequences and the selective regime acting upon a natural population will determine whether these genes are transient, neutral byproducts of mobile extra chromosomal elements or confer adaptive advantage. Ultimately, determining the impact of gene transfer on microbial evolution and the adaptive consequences of horizontally transferred genes will depend upon placing these genomes into well-defined environmental context.

Highly structured biogeographic populations provide a unique biological framework in which to place genome dynamics in geological and evolutionary context. For example, high-resolution multilocus sequence analysis of 130 *Sulfolobus islandicus* strains cultured from geothermal regions in North America, the Kamchatka peninsula, and Iceland revealed that *S. islandicus* is the dominant cultivable *Sulfolobus* species in the Northern Hemisphere and that at least five endemic populations of *Sulfolobus* are isolated from one another by geographic distance. In this highly structured system, evolutionary events are likely to have occurred in each of five geothermal communities independently, and may be correlated with the unique geologic history of their individual locale. Genome sequencing of strains isolated from these endemic populations will allow quantification of rates of horizontal gene transfer relative to nucleotide divergence between isolated populations, and calibration of rates of HGT with the geologic history of isolated geothermal sites. In addition, this novel system may allow the correlation of differences in gene content among closely related individuals to the selective regime of each unique geothermal environment.

* Presenting author

Section 3
Protein Production and Characterization

21
Genes to Proteins: Elucidating the Bottlenecks in Protein Production

M. Hadi[1]* (MZhadi@sandia.gov), K. Sale[1], J. Kaiser[1], J. Dibble[1], D. Pelletier[2], J. Zhou[2], B. Segelke[3], M. Coleman[3], and L. Napolitano[1]

[1]Sandia National Laboratories, Livermore, CA; [2]Oak Ridge National Laboratory, Oak Ridge, TN; and [3]Lawrence Livermore National Laboratory, Livermore, CA

The ability to produce proteins is currently a major biological, physical, and computational challenge in protein research. Given a standard set of conditions, less than 30% of any given genome is expressible in a recombinant host. Protein expression requires complex, lengthy procedures, and specific proteins commonly require individual strategies for optimal expression. Standard bench-level procedures for protein production (expression and purification) do not exist. This lack of validated processes leads to a lengthy search for correct vector, host, expression and purification conditions to yield protein in milligram amounts.

A recent paradigm shift in life science research is from characterizing single genes or proteins (over an investigator's career) to studying whole genomes, proteomes or specific pathways in single "experiments" in a few months. It has been known for some time that while it is quite straightforward to clone and over-express a protein of interest to visualize it on an SDS-PAGE gel. However, it is a completely different matter to obtain purified protein in amounts sufficient for structural and functional studies.

At SNL we have developed successful recombination-based and directed cloning methods for generating large numbers of expression constructs for protein expression and purification. This medium-throughput pipeline is now being automated. This pipeline was initially setup to process targets that included the stress response genes from *D. vulgaris* (GTL funded project). We are using this pipeline to express several hundred ORFs from microbes of DOE relevance (*D. vulgaris, S. oneidensis and R. paulustris*). Each ORF is being expressed using in vivo (*E. coli*) and three in vitro systems (two different prokaryotic and one eukaryotic system). Using the data generated from our expression runs, as well as data generated by other researchers, we are developing computational tools to predict optimal expression system and conditions based on primary sequence and other known properties of the protein. Initially, the prediction software will be specific for prokaryotic systems. However, it will be designed to easily generalize to mammalian systems once training data for mammalian systems is acquired.

* Presenting author

22

High-Throughput Production and Analyses of Purified Proteins

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

[1]Brookhaven National Laboratory, Upton, NY and [2]East Carolina University, Greenville, NC

This work is aimed at improving the efficiency of high-throughput protein production from cloned coding sequences and developing a capacity for high-throughput biophysical characterization of the proteins obtained. Proteins of *Ralstonia metallidurans*, a bacterium that tolerates high concentrations of heavy metals and has potential for bioremediation, are being produced to test and improve the efficiency of protein production in the T7 expression system in *Escherichia coli*. Auto-induction greatly simplifies protein production, as cultures can simply be inoculated and grown to saturation without the need to monitor culture growth and add inducer at the proper time. New vectors allow maintenance and expression of coding sequence for proteins that are highly toxic to the host. Vectors having a range of expression levels have also been made and will be tested for whether tuning the expression level can improve the production of soluble, well folded proteins.

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. As high-throughput production of purified proteins becomes implemented in GTL projects and facilities, reliable analyses of the state of purified proteins will become increasingly important for quality assurance and to contribute functional information. Beam lines at the National Synchrotron Light Source analyze proteins by small-angle X-ray scattering (SAXS) to determine size and shape, X-ray fluorescence microprobe to identify bound metals, and Fourier transform infrared (FTIR) and UV circular dichroism (CD) spectroscopy to assess secondary structure and possible intermolecular orientation. A liquid-handling robot for automated loading of samples from 96-well plates for analysis at each of these stations has been built and implemented with purified proteins. These data will be used as a training set for multivariate analysis of new proteins, to determine whether they are folded properly, obtain information on dynamics and stability, and provide an approximate structure classification. Work has also begun on an automated data reduction and analysis pipeline to process the biophysical information obtained for each protein, and an associated database with a web interface. When fully functional, the system will be capable of high-throughput analyses of size, shape, secondary structure and metal content of purified proteins.

This project is supported by the Office of Biological and Environmental Research of the Department of Energy. Work on auto-induction and vector development also receives support from the Protein Structure Initiative of the National Institute of General Medical Sciences of NIH, as part of the New York Structural Genomics Research Consortium.

* Presenting author

23

A Combined Informatics and Experimental Strategy for Improving Protein Expression

Osnat Herzberg, **John Moult*** (moult@umbi.umd.edu), Fred Schwarz, and Harold Smith

Center for Advanced Research in Biotechnology, Rockville, MD

Improved success rates for recombinant protein expression are critical to many aspects of the Genomes to Life program. This project is focused on determining which factors determine whether or not soluble protein is produced in *E. coli*, and using the results to develop a set informatics and experimental strategies for improving impression results. A three pronged strategy is used: experimental determination of the stability and folding properties of insoluble versus soluble expressers, examination of the cellular response to soluble and insoluble expressers, and informatics and computer modeling.

Informatics methods have now been used to examine a wide range of factors potentially affecting soluble expression, including protein family size, native expression level, low complexity sequence, open reading frame validity, amyloid propensity and inherent disorder. Of these, the most significant ones affecting expression outcome are native expression level, family size, and inherent disorder. The relevance of disorder is being investigated further.

In the first year of the project, we established that the expression of a set of host proteins are consistently upregulated under insoluble protein production conditions, Building on that finding, we are developing reporter constructs utilizing the genes for green fluorescent protein and luciferase, both of which can be assayed by spectrophotometer in intact cells. These reporters will be used to screen a large variety of growth conditions for improved expression of various insoluble recombinant proteins.

Protein stability measurements on a set of proteins using differential scanning calorimetry have been extended, using chemical denaturation with guanidine hydrochoride. So far, the new data support the earlier suggestion that stability is not major factor in determining soluble expression.

This project is supported by Genomes to Life award DE-FG02-04ER63787.

* Presenting author

24

High-Throughput Optimization of Heterologous Proteome Expression

Robert Balint (rfbalint@sbcglobal.net) and Xiaoli Chen

CytoDesign, Inc., Mountain View, CA

High-throughput genomic sequencing has begun to reveal the universe of microbial metabolic capabilities, which can be harnessed to provide new renewable sources of energy, bioremediation, and control of carbon cycling and sequestration, as well as new tools for pharmaceutical discovery, biomolecular synthesis and production, and industrial processes, among others. However, realizing these potentialities will require exhaustive structural and functional characterization of thousands of proteome constituents. Conventional methods for these tasks are so time- and labor-intensive that many years may be required for comprehensive characterization of proteomes of interest, and the costs may be prohibitive. Thus, new methods are urgently needed to accelerate these processes. For structure/function studies *in vivo* and *in vitro*, proteome constituents must be expressed in one or more suitable hosts, accumulating to functional levels in soluble form in their native conformations. They must be purified in sufficient quantities for structural determinations, and also for isolation of analytical reagents such as antibodies. However, many proteome members fail to meet these requirements for one or more of four main reasons: (1) failure to fold before aggregation, (2) failure to fold before proteolytic turnover, (3) instability in the host of choice, or (4) toxicity to the host of choice. An even larger percentage fail to express well enough to provide sufficient material for one or more essential analytical procedures.

The present project addresses the proteome expression problem by adapting selectable reporter systems in *E. coli* for use in high-throughput optimization of heterologous protein expression using peptide chaperones, *i.e.*, small peptides genetically linked to the amino or carboxyl termini of proteome members, which are selected from small libraries for their ability to chaperone folding into native conformations to permit high levels of soluble expression. The chaperone selection systems are based on quantitative coupling of soluble expression to the activity of a reporter which confers a selectable phenotype on the cells. Systems are available for optimization of both secretory and cytoplasmic expression, and the systems permit the optimization of multiple proteins (tens to hundreds) simultaneously.

For convenience and efficiency, proteomes of interest are being optimized for high-level secretory expression in the *E. coli* periplasm. High-level periplasmic expression allows simple recovery and rapid, efficient one-step affinity purification from growth media and/or osmotic shock extracts of continuous or fed-batch cultures, which can be readily scaled to >100 mg/L yields. Periplasmic expression also allows for direct high-throughput selection of antibody affinity reagents for proteome proteins without the need for prior protein purification. The selection system of choice for high-throughput optimization of periplasmic expression is illustrated in the figure below. In this system the reporter is an unstable circular permutation of β-lactamase (β-lacCP), in which a flexible polypeptide linker has been inserted between the native termini, and new "break-point" termini have been introduced at a site within a loop on the surface of the enzyme. The β-lacCP can be fully activated by adding cysteine residues (SH) to the break-point termini, which allow formation of a disulfide bond (SS) to stabilize the locus around the break-point. However, to couple β-lacCP activation to proteome protein expression, a leucine zipper docking mechanism was introduced, in which one cysteine on the CP is replaced by one helix of the zipper, and the other zipper helix, bearing the second cysteine, is linked to a proteome protein of interest (Proteomer). When the

* Presenting author

Proteomer and the β-lacCP are co-expressed in the periplasm of the same cells, formation of the leucine zipper allows formation of the β-lacCP-activating disulfide in proportion to the expression level of the Proteomer. When the Proteomer is equipped with a random sequence library of typically 3-9 residues at its N-or C-terminus, some peptides will (1) protect the Proteomer from aggregation, (2) accelerate folding, (3) facilitate translocation, and/or (4) stabilize the Proteomer, so that the cells expressing these peptides will grow and can be isolated on otherwise non-permissive antibiotic concentrations.

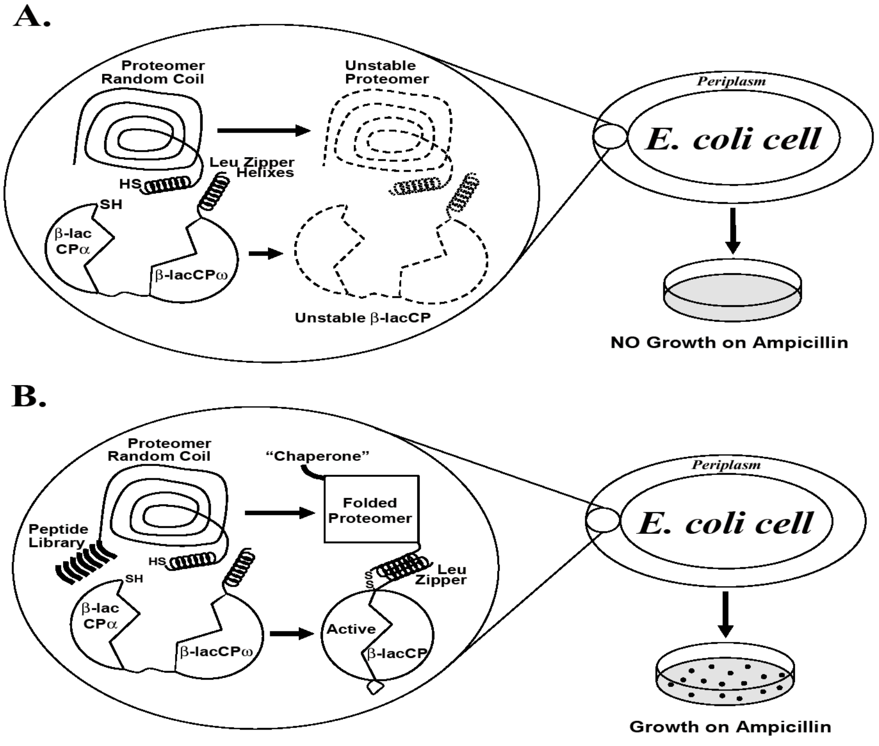


Figure 1. The system is being used to optimize the expression of the proteome of *Pseudomonas putida*, one of nature's most versatile microbes. *P. putida* has the most genes of any known species for breaking down aromatic hydrocarbons, like TNT, and it also reduces a broad spectrum of toxic metals. It encodes at least 80 oxidative reductases for biomass decomposition, and has hundreds of genes for sensing chemicals in the environment. So far, 13 open reading frames (ORFs) selected at random have been expressed in the system with and without N-terminal hexapeptide libraries. Non-permissive antibiotic concentrations were determined for each ORF, and the ORFs were pooled into 3 groups for peptide chaperone selection. At least one peptide was obtained for each ORF, which substantially increased expression, in some cases to >100 mg/L. An added feature of the system is that a suppressible stop codon can be inserted between ORF and leucine zipper helix, so that selections can be performed in a suppressor host, and then selected ORF-peptide constructs can be retransformed into a non-suppressing host for expression without the leucine zipper helix without the need for subcloning.

25

Development of Genome-Scale Expression Methods

Sarah Giuliani, Elizabeth Landorf, Terese Peppler, Yuri Londer, Lynda Dieckman, and **Frank Collart*** (fcollart@anl.gov)

Argonne National Laboratory, Argonne, IL

We are developing novel cellular and cell-free technologies to optimize the expression of cytoplasmic, periplasmic/secreted proteins and protein domains targets from prokaryotic and eukaryotic organism of programmatic interest. The program incorporates technology development and production components and focuses on the use of automated systems and implementation of robotic methods (96-well plate based) to expand the boundaries of current high throughput technology. The technology development aspect applies various expression strategies to target groups and documents the success rate for production of clones validated for expression of a soluble protein target. One example of this component is a project to document expression/solubility outcomes for targets directed to the cytoplasm or periplasmic compartment of *Escherichia coli.* The primary target groups being evaluated in the initial round include the set of periplasmic proteins from *Shewanella oneidensis* as well as a set of simple architecture membrane proteins from *Geobacter sulfurreducens.* The simple architecture membrane proteins contain a membrane anchor or a predicted single membrane spanning helix. For expression/solubility screening of these targets, individual domains adjacent to the membrane region are cloned and targeted to the cytoplasmic or periplasmic compartments.

An important aspect of the production component is the transfer of data and physical resources to experimental labs. The development of target lists and the preferred resources for production of the targets are linked to GTL collaborators with their input and participation solicited at multiple levels including project design, workflow management, and resource distribution. This process involves the generation of specialized vectors to meet experimental requirements [e.g. tag vectors enabling pulldowns for interaction screening or activity screening, or in vivo interaction screening]. One outcome of these efforts is the generation of an expression clone resource for protein targets from various organisms. Our current *E. coli* expression clone library contains over a thousand clones that have been screened for expression and solubility with more than 600 clones that have been validated for successful expression of a soluble protein. The current list of expression clone resource available for distribution can be found on the project website (http://www.bio.anl.gov/combinatorialbiology/GTL.htm). We have extended the scope of available resources based on the need for purified and characterized proteins. In a pilot project to survey requirements for systematic production of proteins, we have purified and distributed >400 purified proteins in mg quantities. Our experience suggests these represent a valuable resource but that considerable effort needs to be expended to define the requirements for large scale protein production.

* Presenting author

26

A Novel Membrane Protein Expression System for Very Large Scale and Economical Production of Membrane Proteins

Hiep-Hoa T. Nguyen[1]* (hiephoa@its.caltech.edu), Sanjay Jayachandran[1], Randall M. Story[1], Sergei Stolyar[3], and Sunney I. Chan[2]

[1]TransMembrane Biosciences, Pasadena, CA; [2]California Institute of Technology, Pasadena, CA; and [3]University of Washington, Seattle, WA

A majority of membrane proteins are very difficult to obtain in any significant quantities, even at milligrams scale since their natural biosynthesis levels often are very low and currently available protein expression systems are not effective for membrane proteins. With the completion of several genome sequencing projects, many large-scale efforts are under way to understand the protein products including the DOE Genome-to-Life project. The lack of effective method for preparative-scale membrane protein synthesis will hamper progress toward a complete understanding of the proteome and prevent us to take full advantage of available sequences, especially of membrane proteins with medicinal importance.

Although currently available *in vivo* protein expression systems are very powerful, capable of producing gram quantities of soluble proteins, their applications to membrane protein synthesis have yielded very poor results. We are working to develop a powerful yet economical host/vector membrane protein expression system utilizing a group of bacteria capable of synthesizing very large quantities of membrane proteins. A series of expression vectors has been created and can be used to express a variety of membrane proteins in these bacteria. Concurrently, other molecular biology tools/protocols are also being developed to genetically engineer these organisms through extensive genome modifications in order to enhance the yield of correctly folded and functional recombinant membrane proteins. Preliminary data from these experiments will be presented.

* Presenting author

27

High-Throughput Methods for Production of Cytochromes *c* from *Shewanella oneidensis*

Yuri Y. Londer, Sarah Giuliani, Elizabeth Landorf, Terese Peppler, and **Frank R. Collart*** (fcollart@anl.gov)

Argonne National Laboratory, Argonne, IL

One of major challenges of post-genomic biology is the development of high-throughput methods for heterologous production of proteins from newly sequenced genomes. The capability for comprehensive production of proteins is an essential component of the systems biology focus of the Genomics:GTL program. Cytochromes *c*, where heme is covalently bound to the polypeptide chain, represent a challenge for heterologous expression systems, since the nascent apoprotein must undergo correct post-translational modification (heme attachment). Our work addresses two major challenges presented by cytochromes *c* with respect to high-throughput production – plate-based detection techniques and the development of a vector resource suitable for periplasmic targeting and efficient expression of cytochromes *c*.

Polyhistidine tags are routinely used for detection of expression and solubility levels in a high-throughput screening methods for protein expression and solubility. However, these tags can interfere with heme attachment and/or folding, especially for multiheme cytochromes, and, therefore, alternative methods to screen for expression and solubility of cytochromes would be useful. We developed a plate-based assay to screen multiple clones for their ability to express *c*-type cytochromes. The assay takes advantage of intrinsic peroxidase activity of heme that can be monitored using commercially available substrates for ELISA. Even though, in the majority of *c*-type cytochromes heme is coordinated by two proteinaceous ligands and not accessible for a molecule of hydrogen peroxide under native conditions, peroxidase activity is preserved upon denaturation of the protein in 6 M guanidine chloride. The method is sensitive enough to detect cytochrome concentrations < 10 µg/ml, which is at least an order of magnitude lower than the concentration of an average recombinant protein in a cell lysate.

The other critical requirement is the availability of high throughput compatible vectors that target nascent polypeptides to the periplasm since targeting to this compartment is necessary for heme attachment to apocytochromes. We designed two families of high throughput compatible vectors that incorporate ligation independent cloning (LIC) sites. One family features a novel LIC-site that allows cloning of targets without the addition of extra residues to the N-terminus (thus preserving a wild-type N-terminus upon the leader peptide cleavage). We have also generated two variants of this vector which enable constitutive co-expression of different periplasmic chaperones. The other family of LIC compatible vectors allows cloning of target genes as fusions to different carrier proteins to facilitate folding and purification. We have used this vector resource to express 30 genes from *Shewanella oneidensis* coding for cytochromes *c* or cytochromes *c*-type domains predicted to have 1-4 hemes. After DNA sequence confirmation, expression and solubility levels were evaluated by SDS-PAGE and the transferred gel pattern stained for heme proteins. Large scale purification and preliminary physico-chemical characterization of individual cytochromes are currently being undertaken.

The submitted manuscript has been created by the University of Chicago as Operator of Argonne National Laboratory ("Argonne") under Contract No. W-31-109-ENG-38 with the U.S. Department of Energy. The U.S. Government retains for itself, and others acting on its behalf, a paid-up, nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.

* Presenting author

28

Towards the Total Chemical Synthesis of Helical Integral Membrane Proteins

Erik C.B. Johnson* (erikj@uchicago.edu) and **Stephen B.H. Kent**

University of Chicago, Chicago, IL

Polytopic helical integral membrane proteins represent an important class of proteins in the cell, yet our understanding of how they function on a molecular level remains elementary due to the inherent difficulties in producing and handling them in their functional forms. ***Chemical protein synthesis*** (CPS) potentially offers an alternative route to the production of integral membrane proteins in quantities sufficient for biophysical studies, yet has been hampered by the difficulty in synthesizing, handling, and purifying peptides that contain transmembrane (TM) domains. These peptides are largely ***insoluble*** in aqueous and mixed organic/aqueous solvents, and show a strong tendency to aggregate. To address this issue, we are developing *reversible backbone protection* as a means to chemically render TM peptides soluble and enable the use of synthetic TM peptides, in conjunction with native chemical ligation, for the total chemical synthesis of helical integral membrane proteins.

29

Selecting Binders Against Specific Post-Translational Modifications: The Sulfotyrosine Example

John Kehoe[3], Jytte Rasmussen[2], Monica Walbolt[2], Carolyn Bertozzi[2], and **Andrew Bradbury**[1]* (amb@lanl.gov)

[1]Los Alamos National Laboratory, Los Alamos, NM; [2]University of California, Berkeley, CA; and [3]Centocor, Horsham, PA

Many cellular activities are controlled by post-translational modifications (PTMs), the study of which is hampered by the lack of specific reagents. The small size and ubiquity of such modifications makes the use of immunization to derive global antibodies able to recognize them independently of context extremely difficult. Here we demonstrate how phage display can be used to generate such specific reagents, using sulfotyrosine as an example. This modification is important in many extracellular protein-protein interaction, including the interaction of some chemokines with their receptors, and HIV infection.

We designed a number of different selection strategies, using peptides containing the sulfotyrosine modification as positive selectors in the presence of an excess of the non-modified peptide as blocking agent. We screened almost eight thousand clones after two or three rounds of selection and identified a single scFv able to recognize tyrosine sulfate in multiple sequence contexts. Further analysis shows that this scFv is also able to recognize naturally sulfated proteins in a sulfation dependent fashion.

* Presenting author

30

Progress on Fluorobodies: Intrinsically Fluorescent Binders Based on GFP

Csaba Kiss, Minghua Dai, Hugh Fisher, Emanuele Pesavento, Nileena Velappan, Leslie Chasteen, and **Andrew Bradbury*** (amb@lanl.gov)

Los Alamos National Laboratory, Los Alamos, NM

Antibodies are the most widely used binding ligands in research. However, they suffer from a number of problems, especially when used in molecular diversity techniques. These include low expression levels, instability and poor cytoplasmic expression, as well the inability to detect binding without the use of secondary reagents. The use of GFP as a scaffold would resolve many of these problems. However, due to the destabilization of GFP folding upon the insertion of extraneous sequences, it has not been possible to use standard GFP as an effective scaffold. Initial attempts to insert diversity into an extremely stable form of GFP (Superfolder GFP) and use phage display were unsuccessful. We have now overcome these problems and have succeeded in selecting GFP based binders which preserve both fluorescence and binding activity. These bind their targets specifically as shown by ELISA, FLISA and flow cytometry, with affinities (measured using surface plasmon resonance) in the nanomolar range.

Fluorescent proteins only become fluorescent when correctly folded. This property becomes extremely useful in the design, selection, screening and use of fluorescent binders, in particular:

- Making libraries; diversity compatible with folding can be selected, screened or monitored
- Monitoring the selection process
- Analyzing expression and affinity of selected fluorescent binders
- Assessing functionality: if it is fluorescent, it is functional
- As a downstream detection signal in e.g. immunofluorescence, FLISA, flow cytometry, biosensors

These binders hold tremendous potential in many different fields, including proteomics and high throughput selection projects, such as the GTL protein production and affinity reagents facility.

* Presenting author

31

High Throughput Screening of Binding Ligands using Flow Cytometry

Peter Pavlik, Joanne Ayriss, Nileena Velappan, and **Andrew Bradbury*** (amb@lanl.gov)

Los Alamos National Laboratory, Los Alamos, NM

Phage display libraries represent a relatively easy way to generate binding ligands against a vast number of different targets. Although in principle, phage display selection should be amenable to automation, this has not yet been described and present selection protocols are far from high throughput. We have examined the selection process in a systematic approach and automated most of the individual steps. Selection is carried out in the microtiter format using 24 targets as the individual selection lot size. We are transitioning from screening by ELISA to using a flow cytometric approach, in which the reactivity of individual antibody clones for numerous different target parameters can be examined simultaneously. This is done by labeling binders fluorescently and coupling analytes to beads which can be distinguished by their intrinsic fluorescence. Presently we label scFvs using a coiled coil approach, in which synthetic fluorescent peptide K coils bind to E coils fused to the scFv. The future use of GFP based binders will eliminate the need for this step. By using specific and non-specific targets, as well as anti-tag antibodies which are able to bind to all binding ligands, we have been able to obtain analyze individual clones from real selections. The information obtained includes binding to specific targets, indications of expression levels, and the degree of polyreactivity/non-specific binding. Depending upon the flow cytometer used, the analysis of each individual affinity reagent clone can be carried out in 1-30 seconds. However, one present difficulty is the export of data, which needs to be carried out manually. With improved data export and analysis, this screening method will be able to handle the throughput desired for the affinity reagents portion of the GTL protein production facility.

* Presenting author

Section 4

Molecular Interactions

32

Exploring the Genome and Proteome of *Desulfitobacterium hafniense* DCB2 for its Protein Complexes Involved in Metal Reduction and Dechlorination

James M. Tiedje[1]* (tiedjej@msu.edu), John Davis[2], Sang-Hoon Kim[1], David Dewitt[1], Christina Harzman[1], Robin Goodwin[1], Curtis Wilkerson[1], Joan Broderick[1,3], and Terence L. Marsh[1]

[1]Michigan State University, East Lansing, MI; [2]Columbus State University, Columbus, GA; and [3]Montana State University, Bozeman, MT

Desulfitobacterium hafniense is an anaerobic, low GC, Gram-positive spore-forming bacterium that shows considerable promise as a bioremediative competent population of sediments. Of particular relevance to its remediation capabilities are metal reduction, which changes the mobility and toxicity of metals, and chlororespiration, the ability to dechlorinate organic xenobiotics. Our work focuses on these two capabilities in *D. hafniense* DCB2 whose genome was recently sequenced by JGI and annotated by ORNL. We are determining the complete metal reducing repertoire of *D. hafniense* and the genes required for respiratory and non-respiratory metal reduction through DNA microarrays produced by ORNL and proteomics. In addition, genomic analysis identified seven putative reductive dehalogenases (RDases) and we are investigating these with genetic and biochemical tools. The status of these efforts is presented below.

Physiology

D. hafniense is capable of reducing iron as well as uranium, selenium, copper and cobalt. With the exception of selenium, the metals were reduced under conditions where the metal was the only available electron acceptor. Selenium was unable to serve in this capacity but was reduced when grown fermentatively. Grown in the presence of Se, the cellular morphology viewed with light microscopy and confirmed with SEM & TEM was elongated with small polyp-like spheres present on the outer surface and in the medium. Energy dispersive spectroscopy (EDS) localized the selenium to the polyps. Staining of the cells with a lipophilic fluorescent dye and osmium profiles from EDS scans suggest that the polyps are surrounded by a membrane. Similar metal localization studies are planned with uranium, copper, iron, and cobalt. Preliminary analysis of gene expression in fermentatively grown cells in the presence vs. the absence of selenium revealed two highly up-regulated genes coding for a radical SAM type protein (Figure 1) and a response regulator consisting of a CheY-like receiver domain and a HTH DNA-binding domain. Expression arrays are underway for identifying genes differentially expressed for the respiratory reduction of iron, uranium, copper and cobalt.

Proteomics

Protein identifications were carried out using 1- and 2-D gel separations of membrane and soluble fractions of *D. halfniense.* Bacteria labeled with ^{15}N are now being grown under fermentative conditions and will be compared to cells grown in ^{14}N in the presence of selenium, uranium, and iron. Sol-

* Presenting author

uble and membrane proteins isolated from mixtures of these *D. halfniense* cultures will be separated by SDS-PAGE and subjected to MS analysis on a ThermoFinnigan LTQ-FT mass spectrometer. The resulting spectra will be analyzed with X! Tandem to identify *D. halfniense* proteins. Comparison of the $^{15}N/^{14}N$ peak ratios are being use to quantify the differential expression of the *D. hafniense* proteins present in the fermentative, selenium-, uranium- and iron-grown cultures, and to correlate these changes with changes in the gene expression profile determine from microarray analysis.

Genetics

Of the seven putative RDase ORFs, two (*rdh1A* and *rdh3A)* were predicted to be nonfunctional due to a nonsense mutation and insertional disruption by a transposase gene, respectively. In each case, the non-functionality was removed through PCR-based procedures resulting in ostensibly full-length RDase genes with 527-aa and 503-aa protein products, respectively, instead of truncated products of 345-aa and 352-aa. Cloning and expression of the altered RDase genes along with neighboring *rdhBC* genes (e.g. *rdh1ABC* and *rdh3ABC*) in *Escherichia coli* resulted in lethality to the host in both cases. Introduction of the *rdh1AB* (without *C*) genes were also apparently lethal. However, *E. coli* clones were obtained for the unaltered *rdh1ABC* cluster and isolated *rdh1C* genes suggesting that expression of the altered *rdh1A* gene is lethal to the *E. coli* host. We are investigating the causes of lethality and the activities of the full-length gene products since this electron accepting pathway may have a central role in the desired and versatile reductive properties of this species.

Biochemistry

The seven RDase genes of *D. hafniense* DCB-2 are being cloned into vectors for expression in *E. coli*. CprA has been cloned into pET44a+ and the resulting plasmid transformed into *E. coli* Rosetta blue cells. Addition of IPTG to the culture at log phase results in significant overexpression of a protein with an apparent molecular mass of approximately 50 kDa. The overexpressed protein has been identified by mass spectral analysis as CprA. It is somewhat soluble under lysis conditions, and has been partially purified with gel filtration chromatography. Both the crude lysate and the partially purified protein exhibit a faint brown coloration, consistent with the presence of an iron-sulfur cluster cofactor.

Based on initial microarray studies, gene 3921 (Contig. 809) is one of only two genes found to be strongly up-regulated under selenate-reducing conditions. This gene encodes a hypothetical protein of 35.8 kDa and comparative sequence analysis suggests the protein likely belongs to the Radical SAM protein superfamily, a group of proteins that utilize iron-sulfur clusters and S-adenosylmethionine to initiate biological radical reactions (Figure 1). Functional studies are currently underway.

Figure 1. Alignment blocks in Radical-SAM proteins. (Ref. Sofia et al., Nucl. Acid. Res. 2001, 29, 1097-1106.)

I. Cluster block (CX $_3$CX $_2$C highly conserved)
consensus: V/I ____G C N__R C __Y C Y __.
gene 3921: IGALNSCPNGCKYCYAN
II. SAM-binding block (Gly-rich, less conserved)
consensus: V_FTGGEPLL .__
gene 3921: IAAKYGIPLQTC
III. Third block (least conserved in superfamily)
consensus: E___LEAIK _L __E _G
gene 3921: ESPLLIGRLKPSDN

* Presenting author

33

Comparison of Conserved Protein Complexes Across Multiple Microbial Species to Evaluate High-Throughput Approaches for Mapping the Microbial Interactome

D. Pelletier[1]* (pelletierda@ornl.gov), G. Hurst[1], S. Kennel[1], L. Foote[1], P. Lankford[1], T. Lu[1], W. McDonald[1], C. McKeown[1], J. Morrell-Falvey[1], D. Schmoyer[1], E. Livesay[3], F. Collart[2], D. Auberry[3], K. Auberry[3], Y. Gorby[3], B. Hooker[3], E. Hill[3], C. Lin[3], P. Moore[3], R. Moore[3], R. Saripalli[3], K. Victry[3], V. Kery[3], S. Wiley[3], and **M. Buchanan**[1]

[1]Oak Ridge National Laboratory, Oak Ridge, TN; [2]Argonne National Laboratory, Chicago, IL; and [3]Pacific Northwest National Laboratory, Richland, WA

The ever-increasing number of available complete genome sequences for bacteria that have been less extensively studied than model systems such as *E. coli* has led to the need for additional methods for system wide functional genomic analysis. Methods (2-hydrid or affinity isolation) for the systematic identification of protein-protein interaction networks have been applied to a number of organisms (e.g., yeast, *C. elegans*, *E. coli*, etc.). It is clear that no single method can be used to fully describe all the interactions in an organism. These data when combined with other genome scale analyses, the transcriptome, the proteome and phenotypic observations, can lead to a better understanding of molecular interactions involved in a biological process.

We have developed a general, scaleable methodology for mapping protein-protein interaction networks, based on affinity isolation of protein complexes, which can be extended to bacterial systems beyond well-characterized model organisms. Specific open reading frames from the annotated genome are identified, cloned and expressed as fusion protein "baits" bearing affinity tags in the target organism (endogenous approach) or in *E. coli* followed by purification then exposed to the proteins from wild type organisms (exogenous approach). Following affinity isolation, protein interactors are identified using mass spectrometry. The metadata are managed using Nautilus LIMS. MS data analyzed by SEQUEST and then undergo computational and statistical analysis. We are using this approach, as part of the Center for Molecular and Cellular Systems, on two phylogenetically distinct bacteria of interest to DOE—the alpha-proteobacterium *Rhodopseudomonas palustris*, and the gamma-proteobacterium *Shewanella oneidensis*. Initial experiments have focused on a number of protein complexes conserved across multiple microbial species to validate this approach for mapping microbial interactomes.

As an example of these studies, the results of mapping the interaction subnetworks for homologs of open reading frames annotated to be part of DNA-directed RNA polymerase, F1F0-ATP synthase and the RNA degradosome will be presented. Results from *R. palustris* and *S. oneidensis* will be compared to results obtained for homologous subnetworks in other organisms that are available in public protein interaction databases (BIND, DIP, etc.). Our data demonstrate the feasibility of our high-throughput approaches for mapping microbial protein interaction networks.

An internal project database and website tracks and summarizes data from all experiments. Publicly accessible views of selected pages of this website show a summary of the status of targeted proteins for the endogenous approach (http://maple.lsd.ornl.gov/cgi-bin/gtl_demo/public_target_status.cgi) and for the exogenous approach (http://maple.lsd.ornl.gov/cgi-bin/gtl_demo/public_ex_target_status.cgi,), and example interactor identifications for a selected subnetwork (http://maple.lsd.ornl.gov/gtl_demo/index.html).

* Presenting author

34

Advanced Technologies for Identifying Protein-Protein Interactions

R. Hettich[1]* (hettichrl@ornl.gov), G. Hurst[1], W. McDonald[1], H. Connelly[1], D. Pelletier[1], C. Pan[1], N. Samatova[1], G. Kora[1], V. Kertesz[1], S. Gaucher[2], T. Iqbal[2], M. Hadi[2], M. Young[2], G. Orr[3], M. Romine[3], D. Panther[3], S. Reed[3], D. Hu[3], E. Livesay[3], B. Hooker[3], S. Wiley[3], S. Kennel[1], and **M. Buchanan**[1]

[1]Oak Ridge National Laboratory, Oak Ridge, TN; [2]Sandia National Laboratories, Livermore, CA; and [3]Pacific Northwest National Laboratory, Richland, WA

The major focus of the **Center for Molecular and Cellular Systems (CMCS)** is the establishment of a *high-throughput identification pipeline for measuring protein-protein interactions*. In the development and operation of this pipeline, several needs have been uncovered that, if addressed, would provide dramatic improvements in overall performance. In particular, improvements are needed in measurement sensitivity, enhanced accuracy, speed of identification, more comprehensive characterizations, and information management tools. The impetus for development of advanced technologies is to alleviate current bottlenecks of the pipeline as well as continue to pursue state-of-the-art technologies for enhanced measurement throughput and data quality. Recent research has focused on key developments in techniques for more comprehensively examining protein interactions, improving LC-MS throughput, applying automated detection of FRET using flow cytometry, and developing computational tools for data mining.

For more extensive characterization of the molecular details of protein interactions, MS methodologies have been developed to extract information from *intact proteins* by either direct top-down MS measurements or by chemical cross-linking of proteins with MS detection of their proteolytic peptides. As a demonstration of the top-down MS approach, tandem affinity purifications were conducted for the Gln family of proteins from *R. palustris*. These proteins are key to sensing internal cellular ammonium levels and transducing the signal. Under ammonium-rich conditions, the proteins GlnK1, GlnK2, and GlnB are unmodified; however, under ammonium-starvation conditions, these proteins are modified by uridylylation and activate the AmtB ammonium transporter and the glutamine synthetase enzyme that are important for nitrogen fixation. Top-down MS measurements provided not only details about the level of uridylyation, but also verified the presence of multiple isoforms of each protein. Chemical crosslinking, used in conjuction with enzymatic digestion and LC/MS/MS, has potential as an important tool for probing protein-protein interactions. Data interpretation is currently a challenge because the "rules" for crosslinked peptide dissociation have not been well studied. We are thus systematically investigating the dissociation pathways open to crosslinked peptides using a series of defined inter- and intra-molecular crosslinked species. Dissociation pathways unique to crosslinked peptides have been found, and are being incorporated into our MS2Assign software. This will allow for more fully automated analysis of crosslinked peptide sequence, a prerequisite for high throughput experiments.

Because of the complexity of protein-protein interactions in microbial systems, it is essential to continue development of methodologies for high throughput measurements. While LC-MS/MS shows tremendous potential for the identification and characterization of protein-protein interactions, the sequential chromatographic separations are either a current or soon to be rate-limiting step for throughput. In order to address this issue, we have undertaken the development of a protocol for fast LC-MS/MS based on a Q-ToF mass spectrometry platform. This setup includes several HPLC columns connected in parallel for multiplexing the runs and solvent washes. By conducting consecutive runs with 30-min. chromatographic separations, we have been able to examine almost

* Presenting author

30 samples in a 24-hour period. The reproducibility and sensitivity are comparable to the longer, more conventional LC-MS/MS methodologies, in which about 12 samples can be analyzed in a 24-hour period. The power of fluorescence resonance energy transfer (FRET) for detecting molecular interactions between proteins can be applied to high throughput screening of protein-protein interactions. Using automated cloning and fluorescence detection approaches, FRET can be used to screen a large number of protein pairs, tagged with fluorescent proteins or fluorescent antibodies. We have combined Gateway compatible cloning with the application of flow cytometry for high throughput FRET analysis in *Shewanella* to leverage existing molecular reagents. Cyan and Yellow fluorescent proteins were successfully expressed in *Shewanella* and were used as donor and acceptor fluorophores in a FRET pair. In addition, we have used FRET analysis between fluorescent antibodies, directed against *Shewanella* periplasmic proteins. We demonstrate that detection of FRET by flowcytometry enables the rapid screening of a large number of cells.

One of the biggest hurdles in the measurements of protein interactions continues to be limitations in computational approaches for data mining. Recent effort has been given to the development of computational methods for integrating top-down and bottom-up MS analyses ("PTMsearchPlus") and deciphering isotopically labeled samples for protein quantification ("ProRata"). In order to automate the interpretation of the top-down MS data, the algorithm "PTMsearchPlus" has been created and tested with several data sets. This program integrates both top-down and bottom-up MS data to exploit the advantages of each approach. The ability to combine high resolution molecular mass data on the intact proteins with the more extensive sequence coverage of the proteolytic peptides from the bottom-up data provides a powerful platform for not only identifying proteins, but also characterizing the presence of post-translational modifications and isoforms. Protein quantification is very important for not only measuring the amount of protein present, but also the elucidation of the stoichiometry of proteins present in complexes. Stable isotope labeling is one of the most widely employed methods for quantification measurements; however, current computational tools to extract the information are quite remedial. We have developed a new tool, "ProRata," which addresses not only the extraction of isotopic information from the labeled samples, but also provides confidence interval scoring for each protein identification.

Research sponsored by the Genomics:GTL program, Office of Biological and Environmental Research, U.S. Department of Energy, under contract No. DE-AC05-00OR22725 with Oak Ridge National Laboratory, managed and operated by UT-Battelle, LLC.

35

Complementary Assays for Validating Protein Interactions Identified by High-Throughput Screening Techniques

Mitchel J. Doktycz[1]* (doktyczmj@ornl.gov), Jennifer L. Morrell-Falvey[1], W. Hayes McDonald[1], Gurusahai Khalsa-Moyers[1], Dale A. Pelletier[1], Stephen J. Kennel[1], Vladimir Kery[2], Galya Orr[2], Dehong Hu[2], Margaret F. Romine[2], David J. Panther[2], Brian S. Hooker[2], H. Steven Wiley[2], and Michelle V. Buchanan[1]

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

Directed experimental validation of protein-protein interactions is essential for confirming associations identified by the Center for Molecular and Cellular Systems' (CMCS) high throughput pipeline and/or predicted by computational techniques. The CMCS core pipeline employs the methods of affinity purification and mass spectroscopy to discover protein interactions in cells. To complement this capability and improve the confidence of protein interaction measurements, the CMCS is also utilizing a combination of live cell and *in vitro* techniques to validate the interaction between specific pairs of proteins. These assays use a common cloning platform that is shared with the core pipeline. The live cell-based approaches include a co-localization assay, fluorescence resonance energy transfer measurements, and yeast two hybrid based tests. Complementing these assays are surface plasmon resonance (SPR) measurements for characterizing protein-protein associations *in vitro*. Use of a suite of technologies will aid in comprehensively identifying protein interactions in a cell. Model complexes are being used to assess the ability of these techniques to characterize transient interactions, low-copy number proteins, or membrane-associated proteins. In addition to confirming physical associations between proteins, these validation assays can provide important functional data on the spatial, temporal, and biophysical context of the interactions. Results from these complementary assays will be shown for protein interactions elucidated from the CMCS core pipeline and/or predicted computationally. Using a combination of methods to assess protein interactions provides the biological community with the high quality data required for understanding the myriad processes that occur in a cell.

* Presenting author

36

Computational Approaches for Aggregating and Scoring Protein-Protein Interaction Data

William R. Cannon[1]* (william.cannon@pnl.gov), W. Hayes McDonald[2], Don S. Daly[1], Denise Schmoyer[2], Gregory B. Hurst[2], Manesh B. Shah[2], Brian S. Hooker[1], Vladimir Kery[1], Stephen J. Kennel[2], H. Steven Wiley[1], and **Michelle V. Buchanan**[2]

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

The GTL CMCS informatics pipeline for *Shewanella oneidensis* MR-1 and *Rhodopseudomonas palustris* protein-protein interaction networks melds information from two sources: data collected internally, including both high-throughput pipeline (see companion abstract: "Comparison of conserved protein complexes across multiple microbial species to evaluate high-throughput approaches for mapping the microbial interactome") and more targeted validation (see companion abstract: "Complementary Assays for Validating Protein Interactions Identified by High-throughput Screening Techniques") assays, and external data sources that both identify interactors and add biological context. A key function of this integrated pipeline is to assess the quality of new data and the processing pipeline itself using: targeted experiments, QC standards, technical and biological replicates, statistical modeling, and statistical process control. Protein interaction networks are then inferred using two different methods. A Bayesian method gives a global analysis of the entirety of the data including false positive and false negative error rates, while a binomial-based maximum likelihood method gives information on error rates of individual bait proteins. This information is captured and published on a web interface (see companion abstract: "The Microbial Interactome Database: An Online System for Identifying Interactions Between Proteins of Microbial Species"). In addition, new decision tools are being developed to automate target selection with the goal of picking the protein targets that will best increase the resolution and breadth of the protein interaction networks.

37

The Microbial Interactome Database: An Online System for Identifying Interactions Between Proteins of Microbial Species

G.B. Hurst[1]* (hurstgb@ornl.gov), D.A. Pelletier[1], D.D. Schmoyer[1], M.B. Shah[1], W.H. McDonald[1], N.E. Baldwin[1], N.F. Samatova[1], A. Gorin[1], B.S. Hooker[2], V. Kery[2], W.R. Cannon[2], D.L. Auberry[2], K.J. Auberry[2], K.D. Victry[2], R. Saripalli[2], H.S. Wiley[2], S.J. Kennel[1], and **M.V. Buchanan**[1]

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

The Center for Molecular and Cellular Systems (CMCS) is generating large amounts of protein-protein interaction data from microbial species through various "pipeline" protocols, as described in other abstracts. To make these data available to the scientific community, we are implementing a Microbial Interactome Database (MID) that will allow users to interact with the data at a variety of levels.

* Presenting author

The primary data source for the MID includes results from mass spectrometric identification of proteins interacting with affinity-tagged "target" or "bait" proteins, resulting from our endogenous and exogenous pipelines. We are currently acquiring these data for *R. palustris* and *S. oneidensis*. As of December 2005, we have performed affinity tagging, isolation, and mass spectrometric analysis for over 210 distinct target proteins ("baits") in *R. palustris,* and 53 proteins in *S. oneidensis*, with multiple replicate analyses in many cases. While rigorous and automated methods for distinguishing authentic interactors from non-specific interactions and background proteins are under development (as described in other abstracts from the CMCS), an empirical method based on filtering at several stages (individual mass spectra, comparison of estimates of interactor protein quantities with an appropriate average value over all comparable samples, frequency of observation, etc.) is currently in place. Interactors identified by this approach include some false positives (artifactual interactors), but also homologs to known complexes and potential novel interactions; experimental, literature-based, or informatic confirmation of these high-throughput results are described in other abstracts. This screen of four interaction subnetworks (DNA-directed RNA polymerase, DNA polymerase, ATP synthase and the degradosome) provides results that are consistent with homologous systems studied by other methods. For example, with α, β, and β' subunits as baits, we redundantly identify the core subunits of the RNA polymerase enzyme (α, β, β' and ω) as well as several sigma factors present only under certain conditions. For DNA polymerase, on the other hand, we have to date identified none of the expected interactors from experiments using β, τ/γ, δ, and χ subunits as baits in *R. palustris*; this result helps define the sensitivity of our current protocol, and suggests that larger culture sizes, more efficient isolation, or more sensitive detection will be required to study subnetworks representing complexes present at the level of ~10 copies per cell.

An internal project database tracks and summarizes data from all experiments, and provides access to CMCS investigators through a web interface. Publicly accessible views of selected pages of this website show a summary of the status of targeted proteins for the endogenous approach (http://maple.lsd.ornl.gov/cgi-bin/gtl_demo/public_target_status.cgi) and for the exogenous approach (http://maple.lsd.ornl.gov/cgi-bin/gtl_demo/public_ex_target_status.cgi,), and example interactor identifications for a selected subnetwork (http://maple.lsd.ornl.gov/gtl_demo/index.html).

We are currently designing a web-based resource to provide access to our results for the scientific community. Important components of this resource include presentation of our complete set of high-throughput "pipeline" results, comparison of putative interactions suggested by pipeline results with results from other databases of interacting protein and literature studies, and incorporation of results from confirmatory experiments such as live-cell imaging and surface plasmon resonance. Critical to the success of this resource will be an interactive filtering scheme to accommodate researchers with diverse scientific requirements, who will require different levels of stringency in the reliability of interactions reported, as well as different tools for extracting subsets of the data. Graphical representations (such as Cytoscape) of protein interaction subnetworks, incorporating various levels of experimental and comparative data, will be included.

Research sponsored by the Genomics:GTL program, Office of Biological and Environmental Research, U.S. Department of Energy, under contract No. DE-AC05-00OR22725 with Oak Ridge National Laboratory, managed and operated by UT-Battelle, LLC.

38

Investigation of Protein-Protein Interactions in the Metal-Reducing Bacterium *Desulfovibrio vulgaris*

Sara Gaucher[1,5]* (spgauch@sandia.gov), Masood Hadi[1,5], Swapnil Chhabra[1,5], Eric Alm[2,5], Grant Zane[3,5], Dominique Joyner[4,5], **Adam Arkin**[4,5], Terry Hazen[4,5], Judy Wall[3,5], and Anup Singh[1,5]

[1]Sandia National Laboratories, Livermore, CA; [2]Massachusetts Institute of Technology, Cambridge, MA; [3]University of Missouri, Columbia, MO; [4]Lawrence Berkeley National Laboratory, Berkeley, CA; and [5]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

Desulfovibrio vulgaris is a sulfate reducing bacteria of interest due to its potential use in bioremediation as well as its economic impact in the petroleum industry (biocorrosion of pumping machinery). This sulfate reducing bacteria has been shown to reduce toxic metals (such as chromium and uranium) to insoluble species making them a good model system for understanding molecular machines involved in bioremediation of contaminated soils and ground water.

Protein complex isolation provides many challenges such as discrimination of non-specific vs. specific complexes and capture of transient partners. We have implemented two contrasting approaches for the isolation of protein complexes from *D. vulgaris*. The endogenous approach involves generating *D. vulgaris* cell lines that produce a "bait" protein of interest fused to an affinity tag (strep tag) that can be captured by chromatography. Lysate from these cells is passed over an avidin column, the bait protein with its associated proteins is captured and can be selectively eluted from contaminating proteins. In the exogenous approach, His-tagged bait proteins from *D. vulgaris* are expressed in *E. coli* cells. These bait proteins are purified to >90% purity and coupled to affinity beads which are then incubated with *D. vulgaris* lysate to capture interacting proteins. The advantage of the exogenous approach is that it is amenable to high throughput and automation. Further, the chance of capturing transient interactions may be increased by varying bait/lysate concentration to drive the equilibrium towards complex formation.

We computationally identified open reading frames (ORFs) that are involved in stress response (including oxygen, heat, pH and salt) based on homology to known stress related genes from other prokaryotic species and have used these ORFs as bait proteins to isolate their protein binding partners. Examples of such bait proteins studied include dnak, ClpX and CooX.. We have also selected proteins that are unique to this sulfate reducer ("signature" genes), expected to yield novel complexes related to sulfate/metal reduction. To validate our methods and address the challenge of non-specific binding, we have included some bait proteins whose interacting partners are well characterized in prokaryotic systems (*E. coli,* -rpoB and rpoC) and have validated our methods by isolating the binding partners of these targets.

In addition to direct complex isolation, we are also measuring protein expression in *D. vulgaris* under different stress conditions using Differential In-Gel Electrophoresis (DIGE) as a supplementary approach to identifying targets for protein-protein interactions studies. Potential binding partners may be gleaned from a list of proteins observed to be co-expressed.

* Presenting author

39

Protein Complex Analysis Project (PCAP): High Throughput Identification and Structural Characterization of Multi-Protein Complexes During Stress Response in *Desulfovibrio vulgaris.*

Project Overview

Dwayne Elias[3], Swapnil Chhabra[1], Hoi-Ying Holman[1], Jay Keasling[1,2], Aindrila Mukhopadhyay[1], Tamas Torok[1], Judy Wall[3], Terry C. Hazen[1], Ming Dong[1], Steven Hall[4], Bing K. Jap[1], Jian Jin[1], Susan Fisher[4], Peter J. Walian[1], H. Ewa Witkowska[4], **Mark D. Biggin**[1]* (mdbiggin@lbl.gov), Manfred Auer[1], Robert M. Glaeser[1], Jitendra Malik[2], Jonathan P. Remis[1], Dieter Typke[1], Kenneth H. Downing[1], Adam P. Arkin[1,2], Steven E. Brenner[1,2], Janet Jacobsen[2], and John-Marc Chandonia[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]University of California, Berkeley, CA; [3]University of Missouri, Columbia, MO; and [4]University of California, San Francisco, CA

The Protein Complex Analysis Project (PCAP) has two major goals: **1.** to develop an integrated set of high throughput pipelines to identify and characterize multi-protein complexes in a microbe more swiftly and comprehensively than currently possible and **2.** to use these pipelines to elucidate and model the protein interaction networks regulating stress responses in *Desulfovibrio vulgaris* with the aim of understanding how this and similar microbes can be used in bioremediation of metal and radionuclides found in U.S. Department of Energy (DOE) contaminated sites.

PCAP builds on the established research and infrastructure of work by another Genomics:GTL project conducted by the Virtual Institute for Microbial Stress and Survival (VIMSS). VIMSS has developed *D. vulgaris* as a model for stress responses and have used gene expression profiling to define specific sets of proteins whose expression changes after application of a stressor. Proteins, however, do not act in isolation. They participate in intricate networks of protein / protein interactions that regulate cellular metabolism. To understand and model how these identified genes affect the organism, therefore, it is essential to establish not only the other proteins that they directly contact, but the full repertoire of protein / protein interactions within the cell. In addition, there may well be genes whose activity is changed in response to stress not by regulating their expression level but by altering the protein partners that they bind, by modifying their structures, or by changing their subcellular locations. There may also be differences in the way proteins within individual cells respond to stress that are not apparent in assays that examine the average change in a population of cells. Therefore, we propose to extend the VIMSS' analysis to characterize the polypeptide composition of as many multi-protein complexes in the cell as possible and determine their stoichiometries, their quaternary structures, and their locations in planktonic cells and in individual cells within biofilms. PCAP will characterize complexes in wild type cells grown under normal conditions and also examine how these complexes are affected in cells perturbed by stress or by mutation of key stress regulatory genes. These data will all be combined with those of the ongoing VIMSS project to understand, from a physical-chemical, control-theoretical, and evolutionary point of view, the role of multi-protein complexes in stress pathways involved in the biogeochemistry of soil microbes under a wide variety of conditions.

Essential to this endeavor will be the development of automated high throughput methods that are robust and allow comprehensive analysis of many protein complexes. Biochemical purification of endogenous complexes and identification by mass spectrometry will be coupled with *in vitro* and *in vivo* EM molecular imaging methods. Because no single method can isolate all complexes, we will

* Presenting author

develop two protein purification pipelines, one the current standard Tandem Affinity Purification approach, the other a novel tagless strategy. Specific variants of each of these will be developed for water soluble and membrane proteins. Our Bioinstrumentation group will develop highly parallel micro scale protein purification and protein sample preparation platforms, and mass spectrometry data analysis will be automated to allow the throughout required. The stoichiometries of the purified complexes will be determined and the quaternary structures of complexes larger than 250 Kd will be solved by single particle cryo EM. An innovative approach to discover weakly interacting protein partners will be investigated that combines rapid freezing of the whole contents of single bacteria, disrupted on the EM grid, and EM. EM tomography of whole cells and of sectioned, stained material will be used to detect complexes in cells and determine their structures at a resolution of 3 nm or better, which is sufficient to recognize known complexes by their characteristic sizes and shapes, to recognize the addition of labile components that may be lost during purification, and to identify instances of previously unknown, candidate complexes that are too labile to be purified by standard methods. New image analysis methods will be applied to speed determination of quaternary structures from EM data. Once key components in the interaction network are defined, to test and validate our pathway models, mutant strains not expressing these genes will be assayed for their ability to survive and respond to stress and for their capacity for bioreduction of DOE important metals and radionuclides.

40

Protein Complex Analysis Project (PCAP): High Throughput Identification and Structural Characterization of Multi-Protein Complexes During Stress Response in *Desulfovibrio vulgaris*

Microbiology Subproject

Terry C. Hazen[1]* (tchazen@lbl.gov), Dwayne Elias[3], Hoi-Ying Holman[1], Jay Keasling[1,2], Aindrila Mukhopadhyay[1], Swapnil Chhabra[1], Tamas Torok[1], Judy Wall[3], and **Mark D. Biggin**[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]University of California, Berkeley, CA; and [3]University of Missouri, Columbia, MO

The Microbiology Subproject of the Protein Complex Analysis Project (PCAP) will provide the relevant field experience to suggest the best direction for fundamental, but DOE relevant research as it relates to bioremediation and natural attenuation of metals and radionuclides at DOE contaminated sites. We will build on techniques and facilities established by the Virtual Institute for Microbial Stress and Survival (VIMSS) for isolating, culturing, and characterizing *Desulfovibrio vulgaris*. The appropriate stressors for study will be identified and, using stress response pathway models from VIMSS, the relevance and feasibility for high throughput protein complex analyses will be assessed. We will also produce all of the genetically engineered strains for PCAP. Two types of strain will be constructed: strains expressing affinity tagged proteins and knock out mutation strains that eliminate expression of a specific gene. Over 300 strains expressing affinity tagged proteins will be produced every year for complex isolation and EM labeling experiments by the other Subprojects. A much smaller number of knockout mutation strains will be produced to determine the effect of eliminating expression of components of putative stress response protein complexes. Both types of engineered strains will be generated using a two-step procedure that first integrates and then cures much of a recombinant DNA from the endogenous chromosomal location of the

* Presenting author

target gene. For this we will develop new counter selective markers for *D. vulgaris.* This procedure will 1) allow multiple mutations to be introduced sequentially, 2) facilitate the construction of in-frame deletions, and 3) prevent polarity effects in operons. The Microbiology Subproject will provide high throughput phenotyping of all engineered strains to determine if any show phenotypic changes. We will test if the tagged proteins remain functional and that they do not significantly affect cell growth or behavior. The knockout mutations will be tested in a comprehensive set of conditions to determine their ability to respond to stress. High throughput optimization of culturing and harvesting of wild type cells and all engineered strains will be used to determine the optimal time points, best culture techniques, and best techniques for harvesting cultures using real-time analyses with synchrotron FTIR spectromicroscopy, and other methods. Finally, we will produce large quantities of cells under different conditions and harvesting techniques for optimal protein complex analyses. To insure the quality and reproducibility of all the biomass for protein complex analyses we will use rigorous QA/QC on all biomass production. We expect to do as many as 10,000 growth curves and 300 phenotype microarrays annually and be producing biomass for 500-1000 strains per year by end of the project. Each biomass production for each strain and each environmental condition will require anywhere from 0.1 – 100 L of culture, and we expect more than 2,000 liters of culture will be prepared and harvested every year. The Microbiology Subproject will optimize phenotyping and biomass production to enable the other Subprojects to complete the protein complex analyses at the highest throughput possible. Once the role of protein complexes has been established in the stress response pathway, we will verify the effect that the stress response has on reduction of metals and radionuclides relevant to DOE.

During the first 3 months, the Microbiology Subproject has supplied several sets of *D. vulgaris* biofilms for EM analysis, three 5 liter cultures of biomass for water-soluble protein complex purification studies, and a 120 liters culture for membrane protein complex purification. We have started using tandem affinity tagging of proteins using three distinct tags in order to purify the protein of interest for detailed characterization. The first uses a "*Strep*-tag" that inserts a streptavidin binding peptide for easy enrichment and we have currently constructed 16 such tags. To attain even higher protein enrichment, however, we are assessing the proven approach of a CTF (a.k.a. SPA) tag that includes a calmodulin binding protein (CBP), a protease (tobacco etch virus) and a 3 x FLAG sites for monoclonal antibody binding versus an "STF" tag that replaces CBP with a streptavidin binding peptide. At issue is the hypothesis that since the latter is only 8 amino acids compared to 125 for CBP, it should be less likely to interfere with localization/orientation of the protein within the cell. All three approaches are currently being assessed with DsrC (DVU2776), a protein in the dissimilatory sulfite reductase pathway that is essential for cell growth via sulfate respiration. Once we have confidently determined the best approach, we intend to tag at least 60 selected proteins in the coming year. Twenty of these proteins will also be tagged with a peptide including a tetracysteine motif that allows *in situ* EM imaging. Since the chemistry upon which this tag is based is thiol chemistry, we must first establish that the sulfide generated by these bacteria does not irrevocably interfere. FtsZ (DVU2499), a cell division protein, is the first candidate for testing the efficacy of this procedure. Within the time scope of this project, we intend to differentially tag >300 of the gene products in *D. vulgaris.* This information is expected lead to a more thorough understanding of not only the proteins involved in metal-reduction but also their protein-protein interactions and characterization of the complete pathway(s) for these activities.

41

Protein Complex Analysis Project (PCAP): High Throughput Identification and Structural Characterization of Multi-Protein Complexes During Stress Response in *Desulfovibrio vulgaris*

Multi-Protein Complex Purification and Identification by Mass Spectrometry

Ming Dong[1], Steven Hall[2], Bing K. Jap[1], Jian Jin[1], Susan Fisher[2], Peter J. Walian[1], H. Ewa Witkowska[2], and **Mark D. Biggin**[1]* (mdbiggin@lbl.gov)

[1]Lawrence Berkeley National Laboratory, Berkeley, CA and [2]University of California, San Francisco, CA

This Subproject of the Protein Complex Analysis Project (PCAP) proposes to test, automate, and use at high throughput multiple methods to purify *in vivo* protein complexes from *D. vulgaris*, identify their polypeptide constituents by mass spectrometry, and determine their stoichiometries. Our goal is to determine an optimum strategy that may include elements of each purification method. These methods will then be used as part of this project's broad goal of modeling stress responses relevant to the detoxification of metal and radionuclide contaminated sites.

Our first purification approach is a novel "tagless" method that fractionates the water soluble protein contents of a bacterium into a large number of fractions, and then identifies the polypeptide composition of a rational sampling of 10,000 – 20,000 of these fractions using MALDI mass spectrometry. Our second purification approach for water soluble proteins is to use and extend the proven Tandem Affinity Purification method (TAP), in which tagged versions of gene products are expressed in vivo and then used to purify the tagged protein together with any other endogenous interacting components. Our third and fourth approaches are specialized variants of the tagless and TAP methods that will be designed to capture membrane protein complexes. A major part of our effort will be the design and construction of automated instruments to speed the throughput of protein purification and sample preparation prior to mass spectrometry, and the development of rapid mass spectrometry data analysis algorithms.

Once established, we will use our optimized methods to catalog as thoroughly as practicable the repertoire of stable heteromeric complexes in wild type cells grown under normal conditions, as well as identify a number of larger homomeric complexes. We will then examine changes in the composition of protein complexes in cells with perturbed stress response pathways. Response pathways will be perturbed either by growing cells in the presence of stressors, including nitrite, sodium chloride, and oxygen, or by mutating cells to delete a component of a stress response pathway. Purified heteromeric and homomeric complexes larger than 250 kD will be provided to the EM Subproject to allow their structures to be determined and any stress induced changes in conformation to be detected. All of these data will be correlated by the Bioinformatics Subproject with computational models of stress response pathways that are currently being established by the Virtual Institute of Microbial Stress and Survival.

Our initial results to date are as follows. We have developed a partially optimized fractionation scheme for the tagless purification strategy and have used it to identify and purify several water soluble protein complexes from *D. vulgaris*. We have also isolated several putative membrane protein complexes following FOS-CHOLINE 12 detergent solubilization of membrane fractions and ion and molecular sieve chromatography. We have identified native gel electrophoresis as having great potential for high throughput, high resolution chromatographic separation of protein complexes

* Presenting author

and are now designing a prototype free flow electrophoresis device to allow use of this separation method within our pipeline. We are developing an iTRAQ based protocol that allows sufficiently accurate quantitation of relative protein levels by LC MALDI TOF/TOF mass spectrometry that co fractionation of polypeptides belonging to known protein complexes can be detected across a series of chromatographic fractions.

42

Protein Complex Analysis Project (PCAP): High Throughput Identification and Structural Characterization of Multi-Protein Complexes During Stress Response in *Desulfovibrio vulgaris*

Data Management and Bioinformatics Subproject

John-Marc Chandonia[1]* (JMChandonia@lbl.gov), Adam P. Arkin[1,2], Steven E. Brenner[1,2], Janet Jacobsen[2], Keith Keller[2], and **Mark D. Biggin[1]**

[1]Lawrence Berkeley National Laboratory, Berkeley, CA and [2]University of California, Berkeley, CA

The Data Management and Bioinformatics component of the Protein Complex Analysis Project (PCAP) has two major goals: **1.** to develop an information management infrastructure that is integrated with databases used by the Virtual Institute for Microbial Stress and Survival (VIMSS) project, and **2.** to analyze data produced by the other PCAP Subprojects together with other information from VIMSS to model stress responses relevant to the use of *D. vulgaris* and similar bacteria for bioremediation of metal and radionuclide contaminated sites.

The high-throughput experiments undertaken by the other PCAP Subprojects will require well engineered databases to capture and store the data in a consistent, structured format that renders it readily amenable to automated analyses. We will develop new infrastructure that is integrated with existing facilities currently deployed by the VIMSS project, in order to leverage existing tools for data analysis between the two projects. This infrastructure will be developed with an eye to scalability, in order to serve additional larger scale projects in the future. Processed data (i.e., complex composition and structures) will be integrated into web pages and other reports accessible through the VIMSS MicrobesOnline website, a public resource for comparative genomics research. Raw data will be made conveniently available to experimentalists for future reprocessing as necessary. We will assess the quality and consistency of all experimental data, as well as determine whether different experimental methods have more or less tendency to recover/detect intact complexes. We will compare our data to other public databases of protein complexes, pathways, and regulatory networks.

In the initial years of the PCAP project, we will prioritize proteins for tagging, TAP, and study by electron microscopy based on analysis of VIMSS data and other bioinformatic predictions. In our first round of prioritization, we have identified 10 *D. vulgaris* proteins as high-priority targets for tagging by the PCAP Microbiology Core, for experimental analysis by TAP/MS and electron microscopy.

All data we obtain on protein interactions will be analyzed in the context of the data currently stored in VIMSS. One of the primary goals of VIMSS is the creation of models of the stress and metal reduction pathways of environmental microbes. Protein-protein interaction data produced by PCAP will be used both to set the structure of these models as well as to parameterize them. Deletion data is particularly useful for inferring the structure of models. For example, if a deletion mutation leads

* Presenting author

to the absence of components from a complex, or the absence of entire complexes, this data may be used to infer details of cellular networks. We also hope to observe entire pathways being up- or down-regulated in response to stress. VIMSS models will be validated against the experimentally observed patterns of complex formation under different stress and deletion conditions.

43

Protein Complex Analysis Project (PCAP): High Throughput Identification and Structural Characterization of Multi-Protein Complexes During Stress Response in *Desulfovibrio vulgaris*

Imaging Multi-Protein Complexes by Electron Microscopy

Kenneth H. Downing[1]* (KHDowning@lbl.gov), Manfred Auer[1], Robert M. Glaeser[1], Jitendra Malik[2], Jonathan P. Remis[1], Dieter Typke[1], and **Mark D. Biggin[1]**

[1]Lawrence Berkeley National Laboratory, Berkeley, CA and [2]University of California, Berkeley, CA

The broad aim of this Subproject of PCAP is to demonstrate the feasibility of using electron microscopy for high-throughput structural characterization of multi-protein complexes in microbes of interest to DOE.

One goal of this work is to characterize the degree of structural homogeneity or diversity of the multi-protein complexes purified by PCAP and to determine the spatial arrangements of individual protein components within the quaternary structure of each such complex. It is already well-established that "single-particle" electron cryo-microscopy has unique capabilities for determining the overall quaternary structure of purified multi-subunit complexes whose molecular weight is greater than ~250 kDa. At a resolution of ~ 2 nm it is possible to locate the positions of individual proteins within such complexes and to then dock previously-determined atomic models of the identified proteins into the envelope of the density map. At resolutions better than 1 nm it is possible to further characterize conformational changes. We aim to increase the throughput of such structure determinations to the level that quaternary structures and docked atomic models are produced within 48 hours of purification of individual, structurally homogeneous complexes.

A second goal is to determine the spatial organization and relative locations of large multi-protein complexes within individual, intact microbes. It has quite recently been established that cryo-EM tomography can be used to produce clearly distinguishable images of larger multi-protein complexes (M_r > ~750 k) within suitably thin, intact cells. Since the cells are imaged in a nearly undisturbed condition, it is possible to count the number of such complexes in each cell as well as to characterize their spatial distribution and their association with other components of subcellular structure. This project will now seek to establish the extent to which it is possible to characterize significant changes in overall subcellular morphology that occur at the single-cell level in response to stress conditions, emphasizing the quantitative changes that occur in the temporal and spatial distributions of various multi-protein complexes. Cryo-tomography will thus be used to translate what is learned from purified multi-protein complexes, isolated from batch cultures, to the more complicated environment of intact cells.

A third goal is to determine whether whole-cell characterization by cryo-tomography can be further supplemented by electron microscopy of cell-envelope fractions and even the whole-cell contents of

* Presenting author

individual, lysed cells. Although both types of material no longer reflect the physiologically native conditions that exist within live microbes, both have a greater capability to gain information about complexes that are either too small to be recognized in tomograms of the molecularly crowded environment within thick specimens, or too labile to remain in tight association over the course of the protocols that must be used for biochemical purification. This approach may reveal an expanded population of multi-protein complexes, some of which may have been previously unknown and others of which may be composed of known, core complexes that still (i.e., immediately after lysis) retain one or more labile components.

Finally, plastic-section electron microscopy is used to translate as much as possible of this basic understanding to the more relevant physiological conditions, both stressed and unstressed, of planktonic and biofilm forms of microbes of interest. Although the morphological recognition of smaller protein complexes is less powerful for this type of electron microscopy, a compensating advantage is that this approach lends itself more easily to labeling – and thus localizing – genetically tagged proteins. Sectioning is also the only technique that can provide images of specimens that are too thick to image as whole-mount materials, while still retaining nanometer resolution. The ultimate goal in using plastic-section microscopy is thus to provide the most complete and accurate information possible about the status of multi-protein complexes, and to do so in a way that can then be used to improve mathematical modeling of cellular responses under the environmental conditions that require bioremediation.

Our initial experiments have been in the areas of single particle cryo EM and EM of biofilms of *D. vulgaris.*

Three separate samples of protein complexes have been evaluated by negative-stain, single-particle electron microscopy. Our pipeline calls for an initial evaluation of each such specimen in uranyl acetate, in neutralized phosphotungstic acid, and in ammonium molybdate, in order minimize misleading characterizations that inevitably occur due to unwanted stain-specimen interactions (e.g. spurious aggregation). One of the specimens appeared to be homogeneous and well dispersed, suitable for taking to the next step of single-particle cryo-EM. The other two specimens, however, appeared to have suffered from having been taken to too high a concentration prior to electron microscopy. From this experience we will adopt a modified protocol in which concentration of protein complexes to levels greater than are actually needed for electron microscopy is carefully avoided.

We have succeeded in growing biofilms of *D. vulgaris* in cellulose dialysis tubing and found the biofilms to cover almost the entire interior of the tube. We have successfully high-pressure frozen and freeze-substituted the biofilms. The sections examined by light microscopy reveal the expected overall biofilm architecture with channel-like areas that are devoid of bacteria or exopolysaccharide (EPS) material. Electron microscopic analysis of biofilm sections reveal loose packing of *D. vulgaris* within the biofilm EPS. Interestingly we found filamentous string-like metal precipitates near the *D. vulgaris*, which may point to structures not unlike the well-characterized *Shewanella* nanowires, which are known to be instrumental in extracellular metal reduction.

44

A Hybrid Cryo-TEM and Cryo-STEM Scheme for High Resolution *in vivo* and *in vitro* Protein Mapping

Huilin Li* (hli@bnl.gov), James Hainfeld, Guiqing Hu, Zhiqiang Chen, Kevin Ryan, and Minghui Hu

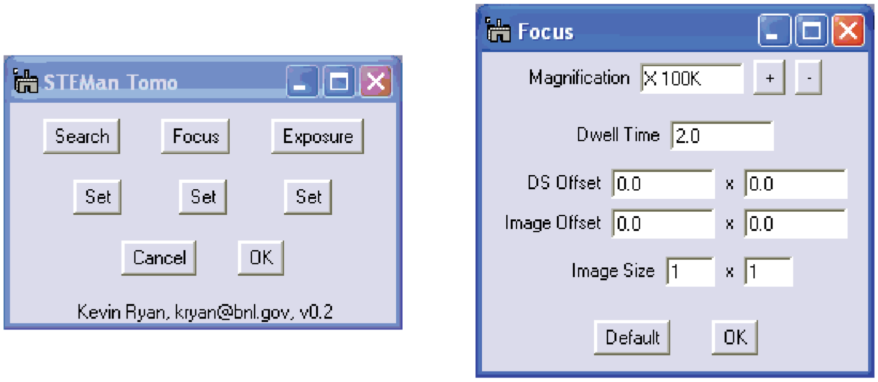
Brookhaven National Laboratory, Upton, NY

The overall structures of biological molecular assemblies and their locations inside a cell are keys to understanding their functions. Our *in situ* hybrid electron tomographic method, which takes advantage of ultra-structural visualization capability of the cryo-TEM and the heavy metal cluster label detection capability of the cryo-STEM, is capable of achieving simultaneously three-dimensional structural visualization and protein mapping. This method can be applied to map protein subunit positions inside a macromolecular assembly, or protein localization inside a microbial cell.

Our approach requires low dose electron imaging and tomographic capability in both TEM and STEM mode. Modern electron microscope comes with these capabilities only in TEM mode, but not in STEM mode. During the past year, we developed a Gatan Digital Micrograph Plug-In (we called it STEMan, see Figure 1) which was based on the Microsoft Visual Studio and Gatan's Software Developer's Kit (SDK). The Plug-In communicates with Jeol FasTEM communication interfaces for microscope setting and with Gatan Microscopy Suite 1.5 for electron beam scanning in STEM mode. With this interface, the damage to the specimen during search and focus can be avoided by offsetting scan and image positions. The process is conceptually similar to the low dose imaging in TEM mode.

Using above-described low dose imaging, we have examined the visibility of 2nm and 5nm unconjugated gold particles embedded in vitreous ice in both TEM and STEM modes. We found that 5 nm particles are visible in both TEM and STEM images, but STEM mode gives better signal to noise ratio. However, the 2nm gold particles in vitreous ice are not visible in TEM mode at low dose and low magnification (25KX) commonly used for tomography. The 2 nm particles are clearly visible in STEM mode at low dose and low magnification (25KX). From these in vitro studies, we expect that the 3 nm functionalized gold particles are detectable in cryo-STEM tomogram of a labeled protein complex either in solution or inside a cell.

Figure 1. STEMan Plug-In graphic user interface. The main menu is on the left, the three top buttons change to the correct settings and then captures the image. The window to the right is displayed when the set button below Focus is pressed. This menu allows the magnification, exposure time, offset, and image size to be set. These values are saved and applied to the microscope when the corresponding image button is pushed.



Kevin Ryan, kryan@bnl.gov, v0.2

Magnification X 100K

Dwell Time 2.0

DS Offset 0.0 x 0.0

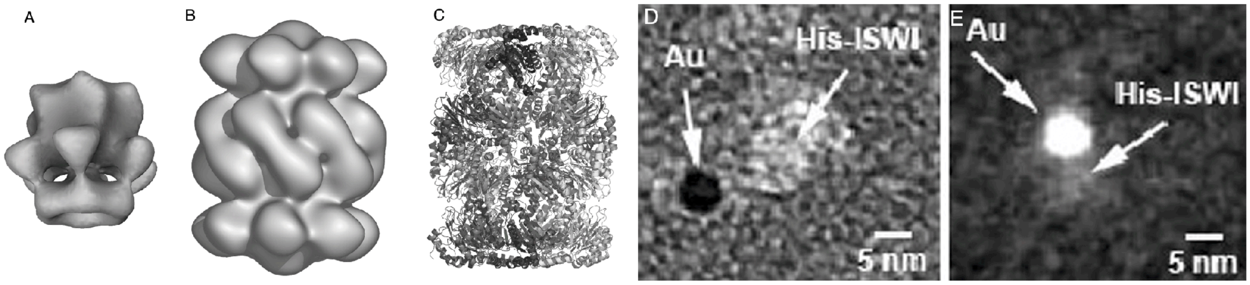
Image Offset 0.0 x 0.0

Image Size 1 x 1

* Presenting author

We have applied EM method to characterizing two microbial protein complexes - the Mycobacterial proteasome and proteasomal ATPase. Both complexes are implicated in resistance to macrophage killing. We found that the proteasomal ATPase is a hexameric structure with a large interior chamber and a relative flat bottom end (Figure 2A)[2]. Cryo-EM reveals a cluster density at the end of the proteasome (Figure 2B) [2,3], not resolved in crystallographic structure (Figure 2C)[3]. The density at the end blocks the entrance of substrate, strongly suggesting that the prokaryotic proteasomes are gated, contrasting to the general belief that the entrance is constitutively open. Our work demonstrates that cryo-EM reveals not only the architecture of microbial macromolecular assemblies, but also their slightly flexible domains that can be crucial to the functions.

Figure 2. Characterization of three macromolecular complexes by TEM and STEM. (A) Mpa, a 500 kDa proteasomal ATPase. (B) A 750 kDa proteasome with closed ends at the top and the bottom. (c) The absence of the end structure in the atomic structure. (D) TEM image of a 6x-histidine tagged protein (ISWI, 140 kDa) labeled with a 5 nm gold particle bearing the Nickel-NTA-group. (E) STEM image of the same sample as in (D).



We have synthesized 1.8 nm, 3 nm, and 5 nm functionalized gold particles. As an example of *in vitro* subunit mapping of multi-subunit complex, the His-tagged DNA remodeling complex is specifically labeled with the 5 nm gold particle with Ni-NTA group (Figure 2D and 2E). Further work will be focused on optimizing labeling condition and application of the technique to more microbial molecular machines both *in vitro* and *in vivo*.

References

1. "Characterization of a *Mycobacterium tuberculosis* proteasomal ATPase homologue," Darwin KH, Lin G, Chen ZQ, Li HL, Nathan, CF. *Molecular Microbiology*, 55, 561–571, (2005).
2. "*Mycobacterium tuberculosis* prcBA Genes Encode a Gated Proteasome with Broad Oligopeptide Specificity," Lin G, Hu GQ, Tsu C, Kunes Y, Li HL, Dick L, Li, P, Chen ZQ, Zwickl P, Weich N, Nathan C. *Molecular Microbiology*, in press.
3. "Structure of the *Mycobacterium tuberculosis* Proteasome and Mechanism of Inhibition by a Peptidyl Boronate," Hu GQ, Lin G, Wang M, Dick L, Xu R, Nathan C, Li HL. *Molecular Microbiology*, in press.

Acknowledgement: This research is supported by a grant from the Office of Biological and Environmental Research of the U.S. Department of Energy (KP1102010).

45

Optical Methods for Characterization of Expression Levels and Protein-Protein Interactions in *Shewanella oneidensis* MR-1

Natalie R. Gassman* (ngassman@chem.ucla.edu), Younggyu Kim, Sam Ho, Nam Ki Lee, Achilles Kapanidis, Xiangxu Kong, Gopal Iyer, and **Shimon Weiss**

University of California, Los Angeles, CA

Unraveling complex biological networks requires robust techniques capable of mapping protein-protein interactions and quantifying gene and protein expression levels. We have been developing a single optical technique, Alternating Laser Excitation (ALEX), which can integrate the analysis of protein-protein interactions and gene expression levels in an accurate, sensitive, and potentially high-throughput manner.

Previously, we reported on the capabilities of ALEX to expand single-pair Förster resonance energy transfer (spFRET) permitting ultra-sensitive analysis of biomolecular interactions by monitoring structure, dynamics, stoichiometries, local environment and molecular interactions. This is accomplished by obtaining D-excitation and A-excitation–based observables for *each single molecule* by rapidly alternating between a D-excitation and an A-excitation laser. 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 (below). Using these ratios, we are now able to examine and characterize protein-protein interactions using both association and conformation information and quantify expression levels using the single-molecule sorting of association state.

We have continued to refine and extend the capabilities of the ALEX methodology, and through the addition of a third laser, 3-color alternating-laser excitation (3c-ALEX), we are now capable of measuring up to 3 intramolecular distances and complex interaction stoichiometries in solution. The 3c-ALEX system substantially extends 2-color ALEX by sorting the molecules in multi-dimensional probe-stoichiometry and FRET-efficiency histograms, and this multiplexing paves the way for advanced analysis of complex mixtures and biomolecular machinery at the single-molecule level. DNA model systems have been used to validate that 3c-ALEX permits FRET-independent analysis of three-component interactions; observation and sorting of singly-, doubly- and triply-labeled molecules present in the same solution; measurements of three intramolecular distances within single molecules, with no requirements for substantial FRET between the probes; and dissection of conformational heterogeneity with improved resolution compared to conventional single-molecule FRET.

To demonstrate the robust nature of these optical techniques and elucidate complex biological networks in a model genome, *Shewanella oneidensis* MR-1, we are applying the ALEX methodology to the regulatory mechanisms governing the transcription of genes in MR-1. Biomolecular interactions involved in signal transduction, formation the RNA polymerase-DNA transcription complex, and in activating or repressing transcription initiation can be monitored and analyzed with these techniques. We have successful reconstituted the transcription machinery from MR-1 and are beginning to examine the relevant protein-protein interactions that induce or repress transcription activation.

* Presenting author

In parallel with our efforts to examine protein-protein interactions in the transcription machinery of MR-1, we are also utilizing the ALEX technique for gene expression analysis. We have utilized DNA model systems to demonstrate the detection and quantification and are currently focusing our efforts at the mRNA level. Our progress in both of these areas will be reported.

46

Protein Interaction Reporter Studies on *Shewanella oneidensis* MR-1

James E. Bruce[1]* (james_bruce@wsu.edu), Xiaoting Tang[1], Wei Yi[1], Gerhard Munske[1], Devi Adhikari[1], Saiful Chowdhury[1], Gordon A. Anderson[2], and Nikola Tolic[2]

[1]Washington State University, Pullman, WA and [2]Pacific Northwest National Laboratory, Richland, WA

We have developed a unique chemical cross-linking system that employs novel compounds that we call "Protein Interaction Reporters" or PIRs that can help identify interactions among proteins in complex biological systems. We have previously applied this strategy to map interactions in a model noncovalent complex [1]. More recently, we have applied the PIR strategy to the microbial system, *Shewanella oneidensis* MR-1, to help elucidate protein interactions that facilitate novel electron transport mechanisms in this system. This report will describe the PIR strategy and highlight our initial PIR results with *S. oneidensis.*

The PIR strategy is based on the use of protein-reactive chemical functionalities that can covalently link interacting proteins in solution, complex mixtures, or within cells. This concept is common to a broad class of cross-linkers that have been exploited for protein structural analysis and limited protein interaction studies. Our PIR strategy combines the utility of chemical cross-linkers described above with mass spectrometry-cleavable features that can help differentiate multiple cross-linking reaction products and facilitate increased proteomics research. For example, the selective cleavage properties of PIR bonds allow release of intact peptides within the mass spectrometer. These peptides can then be studied independently with tandem MS and/or accurate mass analysis to produce data that allows protein identification. In doing so, interactions among proteins can be identified through multiple protein identities resultant from the peptides released from the PIRs. Additionally, the release of intact peptides masses allows differentiation of cross-linking reaction products such as dead-end, intra- and inter-molecular cross-linked species to be identified due to the mathematical relationships that exist between the precursor ions and observed masses that are released from the PIR-labeled species. Finally, the observation of the expected PIR reporter mass also facilitates internal calibration of tandem MS spectra, resulting in improved mass accuracy for tandem mass spectra and improved protein identification capabilities.

The PIR strategy was applied to intact, on-cell labeling studies with *S. oneidensis* following cell culture, harvesting and washing of cells. The current protocols developed for PIR studies utilize two stages of analysis to increase protein identification capabilities. Stage one is carried out with affinity capture of proteins from labeled cells, followed by digestion and tandem MS analysis. This results in a database of PIR labeled proteins. Sites of PIR incorporation are then determined through stage two analysis which involves digestion of all proteins after on-cell labeling, followed by affinity capture of the labeled peptides. These species are then subjected to multiplex LC-FTICR-MS experiments to measure both the intact PIR-labeled peptide masses, and the masses of the species released upon PIR activation. In summary, over 300 proteins were identified though the labeling experiments, many of which were found to be membrane or membrane associated proteins. Several

* Presenting author

of these proteins are known to be critical for electron transport in this system. Additionally, many proteins were found to be plasma proteins and involved in protein synthesis and transport. Based on these initial results, we also carried out electron microscopy cell imaging experiments combined with PIR labeling to visualize sites of PIR incorporation on cell surfaces and interior to the cell plasma membrane. This presentation will highlight the sites of PIR localization, the proteins that were identified through the PIR approach, and discuss the potential for the future of these studies.

Reference

1. Tang, X., Munske, G.R., Siems, W.F. & Bruce, J.E. "Mass spectrometry identifiable cross-linking strategy for studying protein-protein interactions," *Anal Chem* **77**, 311-318 (2005).

47

Implementation of a Data Management and Analysis System In Support of Protein-Protein Interaction Studies of *Shewanella oneidensis* MR-1

Nikola Tolic[1]* (Nikola.Tolic@pnl.gov), Shaun O'Leary[1], Bryce Kaspar[1], Elena Peterson[1], Gunnar Skulason[2], Roger Crawford[2], Gordon Anderson[1], and **James Bruce**[2]

[1]Pacific Northwest National Laboratory, Richland, WA and [2]Washington State University, Pullman, WA

To address the data management and analysis needs of studying protein-protein interactions using cross-linkers and mass spectrometry we deployed the Data Management System (DMS). DMS is a part of the Proteomics Research Information Storage and Management System (PRISM) developed at Pacific Northwest National Laboratory (PNNL) for managing data at the Environmental Molecular Science Laboratory (EMSL) Proteomics Facility[1]. An FTICR mass spectrometer potentially produces hundreds of gigabytes of data daily so a key requirement for a data management system was to provide an affordable solution for data storage, archiving and backup. Since the Proteomics Facility in EMSL handles the same type of data on a much larger scale deploying DMS at Washington State University (WSU) to support their research efforts was a natural and logical selection. It also presented the opportunity to demonstrate DMS's flexibility and scalability for implementation in different environments.

Some key hurdles had to be overcome to deploy the custom built DMS solution to an outside facility. DMS was built to take advantage of PNNL's existing architecture and not necessarily for deployment elsewhere. Its security model is based on our internal network security and it makes extensive use of EMSL's multi-tera byte data archival system. WSU's infrastructure was configured to match PNNL's where DMS required the WSU system to connect to the EMSL archive securely. Because EMSL and PNNL provide a high level of data security collaborating with WSU electronically was a difficult technical and procedural task. The EMSL High Performance Computing & Networking Services group created new enabling technologies including new networks and new software to provide the secure access that was required.

As part of the protein-protein interactions research, the WSU team created a unique strategy they call Protein Interaction Reporter (PIR) which uses specially crafted cross-linker molecules[2] (see related abstract: "Protein Interaction Reporter Studies on *Shewanella Oneidensis* MR-1"). This approach requires development of software tools to analyze and transform the data to relevant information. Raw FTMS spectra are de-isotoped and interpreted using ICR-2LS[3] which was built

* Presenting author

in the EMSL and is integrated with DMS as an automated analysis. ICR-2LS is also used in combination with the commercial software tool called Mascot[4] for peptide and protein identification.

Another tool built specifically for this project is called *XLinks*. This is a set of tools built by at EMSL which employs custom macro functions incorporated into a Microsoft Excel™ template. *Xlinks* is used to extract and report putative cross-linked peptide products from the LC MS datasets by combining precursor and PIR fragmentation data to locate mathematical relationships inherent in data from PIR-labeled peptides. These tools allow automated assignment of PIR-labeled products from multiplexed LC-FTICR-MS datasets. Additional functions were developed and incorporated into *Xlinks* to allow quality control assessment and instrument performance validation on datasets of peptides of protein standards.

The deployment of this data management system to the WSU team has greatly enhanced their ability to continue research in peptide-peptide interactions. Because of the large amounts of data that is generated by the mass spectrometry manual analysis and interpretation becomes prohibitive. Additionally, automated analysis, such as the case with *Xlinks*, is not possible if data files are not stored in a manner that can be effectively accessed, processed, and analyzed. The DMS system at PNNL was an optimal solution to this process had never been deployed outside the PNNL campus. This report highlights the initial application of these critical proteomics tools to academic researchers, and demonstrates the benefits that can be gained by making these valuable National Lab Resources available to the rest of the scientific community. Also discussed in this presentation are the solutions that were devised to the technical and security hurdles that inhibited deployment of DMS to WSU researchers. These issues were critical to the successful implementation of DMS at WSU and the devised solutions are key to doing better science at WSU and other institutions that may implement DMS.

This research was supported by Office of Science (BER), U.S. Department of Energy, Grant No. DE-FG02-04ER63924.

References

1. Kiebel GR, Auberry K, Jaitly N, Clark DA, Monroe ME, Peterson ES, Tolic N, Anderson GA, Smith RD, "PRISM: A Data Management System for High-Throughput Proteomics," *Proteomics*, 2006. In Press.
2. Tang X, M.G., Siems WF, Bruce JE, "Mass spectrometry identifiable cross-linking strategy for studying protein-protein interactions," *Analytical Chemistry*, 2005. 1: p. 311-318.
3. ICR-2LS, http://ncrr.pnl.gov.
4. Perkins, D., D. Pappin, and et al., "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis*, 1999. 20(18): p. 3551-3567.

48

Cell-Permeable Multiuse Affinity Probes (MAPs) and Their Application to Identify Environmentally Mediated Changes in RNA Polymerase and Metal Reducing Protein Complexes

M. Uljana Mayer, Baowei Chen, Haishi Cao, Seema Verma, Ting Wang, Ping Yan, Yijia Xiong, Liang Shi, and **Thomas C. Squier*** (thomas.squier@pnl.gov)

Pacific Northwest National Laboratory, Richland, WA

New generation cell-permeable multiuse affinity probes (MAPs) and complementary protein encoded tags have been developed and applied to identify the regulation of functional interactions between protein subunits in two key multiprotein complexes of *S. oneidensis*; i.e., RNA polymerase and metal reducing complex. Building upon a biarsenical scaffold, distinct reagents with differing colors have been constructed that have orthogonal sequence specificities, permitting their parallel application for high-throughput complex identification. Photoactivatable crosslinking moieties appended to MAPs provide a means to stabilize low-affinity binding proteins prior to cell lysis, providing the first robust means to mediate *in-vivo* crosslinking. An important advantage of this strategy is that a small and nonperturbing tag can be sequentially used to 1) measure the size of protein complexes and their binding affinities, 2) stabilize and isolate intact protein complexes for identification and structural analysis, 3) visualize the location and abundance of the protein complex within individual cells, and 4) the identification of protein function through the light-mediated inactivation of the tagged protein.

Specifically, we have shown:

1. To measure the size of protein complexes in the presence of other cellular proteins, we have taken advantage of the selectivity of MAPs to label tagged proteins, permitting the use of fluorescence correlation spectroscopy (FCS) to measure the size of the diffusing complex. Complementary measurements permit the titration of the protein complex, permitting the determination of the binding affinities of individual subunits.
2. Protein complexes were isolated following the immobilization of MAPs on glass supports, permitting the affinity isolation of tagged proteins and their binding partners. Release of the complex for analysis is facilitated by the mild reducing conditions associated with the release of the complex prior to mass spectrometric analysis. Stabilization of low affinity binding partners is facilitated by the application of newly developed cell permeable MAPs with photoactivatable moieties that permit *in vivo* cross-linking of binding partners.
3. Expression levels of tagged proteins can be directly visualized in live cells or following cell disruption and the separation of proteins on SDS-PAGE. Critical to this capability was the development of a new probe with increased polarity and charge that minimize nonspecific associations associated with high-background fluorescence common to commercially available dyes.
4. Targeted protein inactivation was demonstrated through light-induced protein inactivation, whereby the generation of singlet oxygen was shown to facilitate the formation of zero-length crosslinking and to selectively oxidize surface exposed methionines that are involved in the formation of protein-protein interfaces.

The combination of these capabilities resulting from a single genetic construct facilitates direct comparisons and provides the foundation for high-throughput measurements for a systems level understanding of cellular function.

* Presenting author

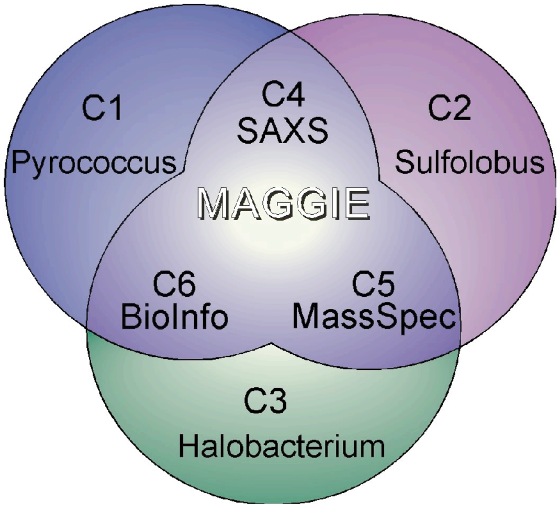
49

Molecular Assemblies, Genes, and Genomics Integrated Efficiently

John Tainer[1]* (jat@scripps.edu), Mike Adams[2], Steve Yannone[1], Nitin S. Baliga[3], Gary Siuzdak[4], and Steve Holbrook[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]University of Georgia, Athens, GA; [3]Institute for Systems Biology, Seattle, WA; and [4]Scripps Research Institute, La Jolla, CA

MAGGIE is developing robust GTL technologies and comprehensive characterizations to efficiently couple gene sequences and genomic analyses with protein interactions and thereby elucidate functional relationships and pathways. The operational principle guiding MAGGIE overall objectives can be succinctly stated: protein functional relationships involve interaction mosaics that self-assemble from independent protein pieces that are tuned by modifications and metabolites. MAGGIE builds strong synergies among the Components to address long term and immediate GTL objectives by combining the advantages of specific microbial systems with those of advanced technologies. The objective for the 5-year MAGGIE GTL Program is therefore to comprehensively characterize the Protein Complexes (PCs) and Modified Proteins (MPs) underlying microbial cell biology. MAGGIE will address immediate GTL missions by accomplishing three specific goals: 1) provide a comprehensive, hierarchical map of prototypical microbial PCs and MPs by combining native biomass and tagged protein characterizations from hyperthermophiles (temperature-trapping otherwise reversible protein interactions) with comprehensive systems biology characterizations of a non-thermophilic model organism, 2) develop and apply advanced mass spectroscopy and SAXS technologies for high throughput characterizations of PCs and MPs, and 3) create and test powerful computational descriptions for protein functional interactions. An overall program goal is to help reduce the immense complexity of protein interactions to interpretable patterns though an interplay among experimental efforts of MAGGIE Program members in molecular biology, biochemistry, biophysics, mathematics, computational science, and informatics. In concert, MAGGIE investigators will characterize microbial metabolic modularity and provide the informed basis to design functional islands suitable to transform microbes for specific DOE missions. These efforts will furthermore test the degree to which metabolic and regulatory pathways can be treated as circuits in which PC and MP components can be swapped in and out to achieve GTL goals in microbial management.



* Presenting author

50

The MAGGIE Project: Identification and Purification of Native and Recombinant Multiprotein Complexes and Modified Proteins from *Pyrococcus furiosus*

Francis E. Jenney Jr.[1]* (fjenney@arches.uga.edu), Farris L. Poole II[1], Angeli Lal Menon[1], Rathinam Viswanathan[1], Greg Hura[2], John A. Tainer[2], Sunia Trauger[2], Gary Siuzdak[2], and **Michael W. W. Adams**[1]

[1]University of Georgia, Athens, GA and [2]Scripps Research Institute, La Jolla, CA

The genes that encode multiprotein complexes (PCs) or post-translationally modified proteins (MPs), such as those that contain metal cofactors, in any organism are largely unknown. We are using non-denaturing separation techniques coupled to mass spectrometry (MS) and metal (ICP-MS) analyses to identify PCs and MPs in the native biomass of a model hyperthermophilic organism of DOE interest, *Pyrococcus furiosus* (Pf), which grows optimally at 100 °C. By analyzing the native proteome at temperatures close to 100 °C below the optimum physiological temperature, we will trap reversible and dynamic complexes thereby enabling their identification and purification. Samples of the more abundant PCs and MPs obtained from native biomass are being used directly for structural characterization by Small Angle X-ray Scattering (SAXS), which provides information on overall mass, stoichiometry of subunits, radius of gyration, electron pair distances and maximum dimension. Recombinant versions of the less abundant PCs and MPs are being obtained by multiple gene expression systems designed using information from native biomass analyses and bioinformatic approaches. Robotic-based expression analyses are being used to assess the production of recombinant PCs and MPs, and these will be produced on a preparative scale for structural characterization. The recombinant aspect of the project utilizes the infrastructure developed for a previous structural genomics effort with Pf. This yielded the stable, recombinant forms of almost 400 Pf proteins. The non-recombinant aspect builds on extensive expertise in purifying and characterizing native multiprotein complexes from Pf using up to 1 kg of biomass as starting material. Results will be presented from complexes purified from Pf grown under at 95°C using peptides as the primary carbon source.

* Presenting author

51

Systems Approach in a Multi-Organism Strategy to Understand Biomolecular Interactions in DOE-Relevant Organisms

Nitin S. Baliga[1] (nbaliga@systemsbiology.org) and John Tainer[2] (JATainer@lbl.gov)

[1]Institute for Systems Biology, Seattle, WA and [2]Lawrence Berkeley National Laboratory, Berkeley, CA

In this component of the MAGGIE project we will apply systems approaches to the study of molecular complexes. The important point to note is that this strategy will result in reduced costs by complementing through a versatile organism such as *Halobacterium* NRC-1 the lack of expensive systems biology tools for other organisms of the consortium. One of the many important contributions of this component will be the Gaggle. The Gaggle is a simple, open-source Java software environment that helps to solve the problem of software and database integration. Guided by the classic software engineering strategy of separation of concerns and a policy of semantic flexibility, it integrates existing popular programs and web resources into a user-friendly, easily-extended environment. We demonstrate that four simple data types (names, matrices, networks, and associative arrays) are sufficient to bring together diverse databases and software.

Section 1

OMICS: Systems Measurements of Plants, Microbes, and Communities

52 MEWG

Meta-Proteomic Study of a Microbial Community Response to Cadmium Exposure

C.M.R. Lacerda[1], L. Choe[2], and **K.F. Reardon**[1]* (reardon@engr.colostate.edu)

[1]Colorado State University, Fort Collins, CO and [2]Cornell University, Ithaca, NY

Microbial communities are the basis for engineered environmental bioprocesses. However, current methods, including 16S rDNA-based techniques, are incapable of providing information on the function of the community or its members, and thus the community remains a black box from a functional perspective. Microbial community proteomics has the potential to detect proteins expressed in an environment under different conditions. In this approach, a mixture of microorganisms can be viewed as a meta-organism, in which population shifts are a form of functional response. Despite its possible advantages, the use of meta-proteomics has been almost completely unexplored.

In this project, meta-proteomics was used as a tool to obtain functional information about the response of a microbial community to cadmium stress. Cadmium (10 mg/L) was added to a mixed culture and protein samples collected after 0.25, 1, 2 and 3 hours. Comparison of the two-dimensional gels from Cd-exposed and control cultures revealed that the community "reacts" by changing its protein profile, with both increased and decreased expression of substantial numbers of proteins that change with time. Within 15 minutes of exposure, nearly 20% of the proteins detected on the 2-D gels were found to change in level by three-fold or more. Mass spectrometric analysis was also performed to identify proteins that play central roles during the cadmium shock. Fifty proteins have been identified to date. Metabolic enzymes make up the largest functional group, followed by proteins involved with defense, protein synthesis and storage, and energy metabolism. A variety of temporal expression patterns was noted, with protein functional groups displaying one or more patterns.

This study demonstrated that proteins can be identified from a community of unsequenced organisms. Our proteomic analysis revealed significant shifts in the community physiology for both short and long term metal exposure, insights that could not have been obtained using traditional 16S rDNA methods. These results clearly demonstrate the ability of the meta-proteomic approach to detect changes in microbial communities and support its use for determining functional states of a microbial community.

* Presenting author

53

Investigating Novel Proteins in Acidophilic Biofilm Communities

Christopher Jeans[1], Clara Chan[2], Mona Hwang[1], Jason Raymond[1], Steven Singer[1], Nathan VerBerkmoes[3], Manesh Shah[3], Robert L. Hettich[3], **Jill F. Banfield**[2], and Michael P. Thelen[1]* (mthelen@llnl.gov)

[1]Lawrence Livermore National Laboratory, Livermore, CA; [2]University of California, Berkeley, CA; and [3]Oak Ridge National Laboratory, Oak Ridge, TN

A large fraction of proteins deduced from microbial genomes do not match any recognized sequences, and there is no systematic method at hand to analyze this collection known as "hypothetical proteins". These have for the most part remained in the hypothetical category because authentication by protein detection or gene expression has been difficult; however, it is likely that a large portion are important, in some cases essential, for microbial fitness under natural conditions. The void in our knowledge regarding this dark side of genomics, representing nearly half of protein sequences, is a formidable obstacle to nearly every investigation enabled by full genome sequence information. In our study of acidophilic biofilm communities, we are focused directly on the problem of hypothetical proteins within a well-defined system in one of the most extreme limits of habitability. These microbes band together in very acidic streams of underground mine tunnels to form floating, matrix-bound, self-sustaining communities that derive electrons from iron (II) while in turn catalyzing the dissolution of pyrite (FeS_2). Such conditions result in the biological generation of acid mine drainage (AMD), ultimately causing many orders of magnitude greater metal dissolution and environmental acidification than abiotic forces alone. Since AMD is so prevalent and potentially disastrous to US inland waters associated with principle energy resources such as coal and uranium, understanding the mechanisms by which biofilm communities generate AMD is a priority with DOE's Office of Biological and Environmental Research. Also relevant to the Genomics: Genomes to Life Program, our project will provide fundamental scientific contributions resulting from a systematic and thorough study of the hypothetical protein problem, including the development of methods to be used in these and other systems where a description of novel proteins will provide the keys to overall biological function.

In addition to the detailed geochemical and microbiological description of AMD biofilms by Jill Banfield and colleagues at UC Berkeley, the foundation of our current investigation is genomic and proteomic datasets that are directly linked for each of several dominant organisms. Characterization of abundant species in one such biofilm and the associated genome analyses resulted in reconstruction of near complete or partial genomes for two different *Leptospirillum* bacteria and three archaea[1], and further work is now extending the sequences within this biofilm and also to ones with differing population structures. In our attempt to understand essential biofilm metabolic activities and the partitioning of functions between individual organisms, proteins native to an environmental biofilm were analyzed by shotgun MS proteomics, carried out by Robert Hettich and associates at ORNL. This approach coupled with protein biochemistry confirmed that many hypothetical proteins are expressed at detectable levels in several biofilm organisms, and also indicated that a majority of abundant extracellular proteins are novel[2]. This groundwork has enabled us to examine the novel proteins expressed in biofilms harvested from several neighboring locations, in carefully measured geochemical conditions, using a combination of computational analyses/predictions, and protein biochemistry coupled with MS identification.

To reveal any functional features of proteins overlooked by the first level of bioinformatics, we have drilled further into protein sequence homology, initially using a database containing the completed

* Presenting author

genome sequences of 250 prokaryotes. We interrogated this data with several hundred novel protein sequences that were detected by MS proteomics in one of the dominant organisms in the AMD biofilms, *Leptospirillum* group II. Using parameters that can identify distant homology, we were able to classify more accurately "unique" and "conserved" protein sequences. Proteins found only in *Lepto.* II, in addition to those similar to (but distinctly different from) proteins in the other abundant biofilm bacterium, *Lepto.* group III (e.g., *L. ferrodiazotrophum* sp. nov.[3]), are considered unique. Genes encoding these proteins may have arisen in direct response to environmental selection, and are likely to be vital to AMD biofilm functions. Also notable are the similarities found between the novel proteins of *Lepto.* II and III to those in "functionally related" organisms, such as *Geobacter metallireducens* and *Ralstonia metallidurans*, both of which can be found in toxic metal-rich niches. Other unknown proteins that are homologous to 20 – 90 other microbial sequences indicates conservation of core functions that remain unidentified. Descriptions of proteins in this latter category will have perhaps the most impact on genome-enabled microbiology. In addition to individual protein sequence homology, analysis of gene synteny in sequences flanking novel genes is a measure of evolutionary conservation. In some cases clues to functions were found from annotations to neighboring genes within homologous operons, indicating a biochemical pathway or protein complex in which a novel protein resides. These analyses are being used to indicate interesting novel protein targets to pursue, using the abundance inferred from proteomics datasets as a guide.

In addition to these predictive criteria, we also use biochemical properties such as protein mass, isoelectric point and some sequence information (e.g., predicted or experimentally determined signal/transit peptides) to facilitate the isolation and identification of proteins for experimental characterization. This is being pursued mostly in biofilm extracts that have been fractionated into extracellular and membrane-associated proteins, both enriched with respect to novel proteins, with a major goal to characterize the functional components of protein complexes. There are several examples to illustrate this approach. From biofilm extracts, we purified one of the most abundant extracellular proteins, identified its *Lepto.* II gene and found sequence similarity with only one unknown protein sequence (from *Ralstonia metallidurans*) in the 250 fully sequenced microbes. This soluble red cytochrome of 16 kDa contains an unusual heme, has a reduction potential of 640 mV – close to that of soluble iron at pH 2.0 – and rapidly oxidizes iron (II to III), yielding a unique absorption peak at 579 nm (ref. 2). We used antibodies to localize this 'cytochrome 579' in biofilms, and found that the protein is highly concentrated near the surface of *Lepto.* II cells. Complementing these results, we have isolated another abundant, novel protein from *Lepto.* II that is primarily membrane associated and has similar spectral characteristics to cytochrome 579, but is a much larger protein. This 57kDa cytochrome forms a complex with cytochrome 579, indicated by the co-electrophoresis of the two proteins in nondenaturing gels. Further testing will establish the role of the larger cytochrome in iron redox reactions, and the function of the two proteins in this complex, towards a more accurate description of the initial steps in electron transfer from iron sulfate in mine solutions to the cellular electron transport system. We have also purified two other protein complexes and several individual proteins from both extracellular and membrane fractions using similar methods, and these are targets for detailed characterization. These several results indicate that our approach can be used to identify novel gene products and biochemical functions.

To process many more proteins in future work, we have started to separate biofilm proteins using a variety of selective chromatographic media and identify the principle proteins in these fractions using LCQ_MS. In conjunction with size exclusion and chromatofocusing, these methods provide us with a powerful approach to isolate many of the novel proteins along with any known proteins associated with them, from both the environmental microbial community and from biofilm isolates cultured in lab. Multiple fractions from these columns will be initially assayed for several key biochemical activities such as hydrolases, oxidoreductases and polysaccharide synthases that are suspected to play important roles in biofilm formation and function. The development of this type of systematic approach for the validation and description of hypothetical proteins, used within a defined biological system, is a test case for the full utility of proteogenomic information.

This Genomics: Genomes to Life project centered at UC Berkeley is performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

References

1. Tyson et al., 2004, *Nature* **428**:37-43
2. Ram et al., 2005, *Science* **308**:1915-20
3. Tyson et al., 2005, *Appl Environ Microbiol* **71**:6319-24

54

Proteomics Measurements of a Natural Microbial Community Reveal Information about Community Structure and Metabolic Potential

Nathan C. VerBerkmoes[1]* (verberkmoesn@ornl.gov), Rachna J. Ram[2], Vincent J. Denef[2], Gene W. Tyson[2], Brett J. Baker[2], Manesh Shah[1], Michael P. Thelen[3], **Jillian Banfield**[2], and Robert L. Hettich[1]

[1]Oak Ridge National Laboratory, Oak Ridge, TN; [2]University of California, Berkeley, CA; and [3]Lawrence Livermore National Laboratory, Livermore, CA

Microbial communities play key roles in the Earth's biogeochemical cycles, but for the most part are very poorly understood. For example, microorganisms from a number of lineages thrive in acid mine drainage (AMD), one of the most extreme environments on Earth. Recently, we have completed a proteogenomic investigation of a natural AMD consortia sample to determine the biochemical basis for adaptation (Ram et al, Science, 2005). The genomic data enabled elucidation of abundant proteins by providing the database necessary for identification of peptides from whole cellular, membrane, and periplasmic/ extracellular fractions of the biofilm samples. The proteomic data provided the first *in situ* analyses of community structure and metabolic potential. MS-based "shotgun" proteomics measurements with multidimensional liquid chromatography – linear trapping quadrupole mass spectrometry provided confident measurement of 2036 proteins from the AMD sample, with identifications corresponding to all five dominant species in the biofilm.

For the complete characterization of the proteome of complex natural microbial communities, we have developed a systematic experimental plan to push the capabilities of current mass spectrometry techniques as well as integrate the next generation of technology. This will involve evaluation and implementation of a higher performance MS technology employing a state-of-the-art LTQ-FT-Orbitrap (Thermo Electron) instrument for deeper and more comprehensive characterization and quantification of the proteins and protein complexes important for community structure. The ultimate goal will be to define and demonstrate experimental protocols to begin to unravel the molecular details of natural microbial communities. Specific tasks are directed at examination of spatially and temporally distinct AMD biofilms, establishment of advanced mass spectrometry methodology for more comprehensive proteomic information, elucidation of the details of microbial strain variant diversity with high mass accuracy and MS^3 experiments, and quantification of protein abundances to understand the major investments of cellular resources.

Recent work has been focused on the proteome characterization of new AMD microbial communities distinct but related to the original AMD sample. These include the AB front, C pink, and UBA Ultraback locations of the mine. These proteomes are currently being analyzed by similar methods

* Presenting author

used in the previous study and over 1,500 proteins have been identified from each biofilm. We are currently comparing these datasets to the original proteome dataset to determine conserved protein expression between all samples and to determine those proteins expressed only in given biofilms.

We are in the process of transitioning from current MS technology to more advanced techniques to provide better depth of coverage, smaller sample size requirements, and to unravel the details of microbial strain variation in AMD communities. The proteomic analyses of previous and current samples generated many mass spectra that could not be assigned to any protein. There are several reasons for this, two of which relate to problems with the genomic database. Peptides from proteins that differ significantly from those encoded in the community genomic dataset are unlikely to be matched. If there are substitutions throughout the protein, matching of all peptides (or all but one peptide) may be precluded. In this case, the protein is unidentifiable. Alternatively, a subset of peptides may be found, resulting in (i) assignment of the peptides to a protein that is not identical to the dominant form in the community and (ii) incorrect evaluation of the relative abundance of that protein. For these reasons, proteomic analyses must move away from composite genome sequence-based analyses to utilize the full database of gene types detected in each community. A second reason for unassigned spectra is the presence of organisms in the community that were not sampled in the genomic dataset (due to different organism membership or incomplete genome coverage). It is important to expand the analytical capabilities of the MS-based proteomics methodology for identifying and obtaining detailed biological information for every protein in an AMD sample. This requires identification of a very large number of proteins and their post translations modifications, truncation products, and associated strain variability. The greatest challenges for analysis of a microbial community are the large number and relatively low concentrations of many proteins within the system. For all initial studies, protein separation were conducted on an integrated 2-dimensional nano columns equipped with nanospray MS. For the next phase of this project, a new hybrid LTQ-FT-Orbitrap will be integrated into the pipeline and be used for all sample characterization. High mass accuracy, high sensitivity and high dynamic range capabilities are features of this new instrument, which can be coupled with a linear ion trap mass spectrometer capable of rapid data-dependent MS/MS and MS^3. This instrument can easily be coupled with multidimensional chromatography. The high mass accuracy will allow for much more confident identifications of MS/MS spectrum due to the accuracy at which the parent mass of the peptide can be determined (<3 ppm). Furthermore, if desired, MS/MS spectra can also be analyzed in the Orbitrap allowing for <3 ppm mass accuracies on fragment ions in MS/MS or MS^3 experiments. This is critical in community samples where databases are larger and sample variability will be much greater. More comprehensive proteomic analysis will be achieved due to the high dynamic range of the instrument, because data-dependent MS/MS events will enable measurement of peptides that cannot be detected in full scan mode on the linear ion trap. The linear ion trap is capable of MS^3 experiments, which, along with high mass accuracy of the intact peptide and MS/MS spectra, will greatly increase the accuracy of *de novo* sequencing techniques outlined below.

To interpret the LC-MS/MS data from the samples, proteome bioinformatics, in particular, Sequest and DBDigger search engines, will be employed and web-based data repositories will be created. One major goal of this subproject is to develop new informatic tools for data analyses, data mining, and data display critical for the challenges associated with these studies. New computational techniques will immediately be implemented into the proteome informatics pipeline. We are currently evaluating a number *de novo* sequencing algorithms for identification of unmatched peptides. The use of high mass accuracy MS/MS spectra will greatly enhance the accuracy of these algorithms for correctly determining peptide sequences directly from MS/MS spectra.

The final task of this work involves absolute and relative quantification of proteins from AMD biofilms. While exact quantification of proteins from simple isolate microbial proteomes is challenging the quantitation of proteins directly from environmental samples provides a much greater diversity of challenges. We are currently investigating absolute quantification of proteins in the AMD community samples with isotopically labeled synthetic peptides from well characterized peptides from the community. The use of isotopic encoded affinity tags (ICAT) and ^{18}O Water labeling will be investigated for relative quantification of proteins between multiple AMD biofilms.

Research sponsored by the Genomics:GTL program, Office of Biological and Environmental Research, U.S. Department of Energy, under contract No. DE-AC05-00OR22725 with Oak Ridge National Laboratory, managed and operated by UT-Battelle, LLC.

55

Computational Algorithms and Software Tools for Quantitative Shotgun Proteomics

C. Pan, B. Zhang, G. Kora, N.C. VerBerkmoes, D. Tabb, W.H. McDonald, D. Pelletier, E. Uberbacher, G. Hurst, R. Hettich, and **N.F. Samatova*** (samatovan@ornl.gov)

Oak Ridge National Laboratory, Oak Ridge, TN

Biological organisms respond to many environmental or physiological stimuli by adjusting the expression levels of proteins. The quantification of protein abundance and detection of differential protein expression under various experimental conditions are fundamental and challenging problems in proteomics. Addressing these problems with classical two-dimensional gel electrophoresis (2DE) has some obvious disadvantages due to its low detection sensitivity and linearity, poor solubility of membrane proteins, limited loading capacity of gradient pH strips, low reproducibility of gels, limited throughput, and small linear range of visualization procedures. Recently, alternative approaches based on either *stable isotope labeling* or *label-free* mass spectrometry (MS) shotgun proteomics have emerged as a high throughput technique for measuring the relative abundance of thousands of proteins from different cell cultures.

Because of highly complex sample handling in proteome measurements (as described in other abstracts from the Center for Molecular and Cellular Systems, CMCS), proteome quantification requires rigorous statistical approaches. To provide robust quantification of relative protein abundance and sensitive detection of biologically significant differential and correlated protein expression, we are developing advanced statistical methods coupled with suitable software tools that are made available to users as open source (please, pick up the distribution CD with the software or email a request).

ProRata for relative quantification of mixed stable-isotope-labeled proteomes

To extract relative peptide and protein amounts from mass spectrometric measurement of mixed stable-isotope-labeled proteomes, we developed a computer program, called ProRata. To improve both quantification accuracy and quantification confidence, we systematically optimized the core analysis steps for robust quantification: chromatographic peaks detection, peptide relative abundance estimation, and protein relative abundance estimation. Our novel parallel paired covariance algorithm has largely enhanced the signal-to-noise ratio of the two isotopologues' chromatograms and, as a result, has enabled much more accurate peak detection. Principal component analysis (PCA) was employed to estimate peptide abundance ratios and demonstrated superior estimation accuracy than the tradi-

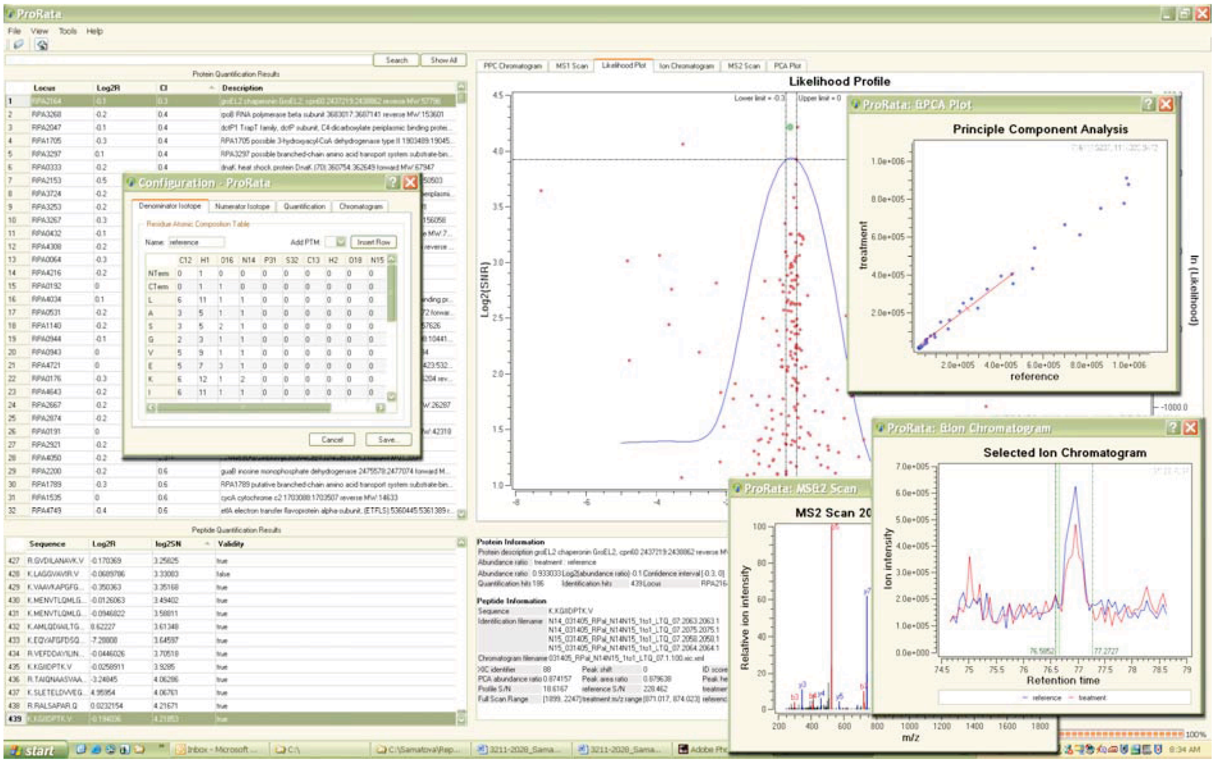
* Presenting author

tional methods based on peak height and peak area. It was observed that the relative quantification of the standard proteome mixtures is of highly variable accuracy for peptides and consequently for proteins. To estimate quantification error, we proposed a novel signal-to-noise measure derived from principal component analysis and showed a linear correlation of this measure with the peptide ratio estimation in the standard mixtures. Finally, maximum likelihood estimation (MLE) was used for protein relative quantification from the PCA-estimated abundance ratios of its proteolytic peptides. MLE not only showed more accurate protein quantification and better coverage than the widely-used RelEx program but also a more robust estimate of a confidence interval for each differential protein expression ratio. For an automated data processing and streamlined data visualization, these algorithms were integrated into a ProRata computer program (see Figure for a sample display).

Detection of differential and correlated protein expression in label-free shotgun proteomics

We performed a systematic analysis of various approaches to quantifying differential protein expression in the eukaryotic *Saccharomyces cerevisae* and prokaryotic *Rhodopseudomonas palustris* LC-MS/MS label-free shotgun proteomic data. First, we showed that, among three sampling statistics, the *spectral count* has the highest technical reproducibility followed by the less-reproducible *peptide count* and relatively non-reproducible *sequence coverage*. Second, we used spectral count statistics to measure differential protein expression using four statistical tests: Fisher's exact test, G test, AC test, and t-test. For the yeast data set with spike proteins, the first three tests performed similarly on a pair-wise comparison of multiple experiments. Their False Discovery Rate (FDR) was less than 0.4% for a 10-fold change and less than 0.7% for a 5-fold change, even with a single replicate. For a 2-fold change,

Figure. The display of ProRata analysis results for ^{14}N:^{15}N labeled *R. palustris* proteome. The ProRata graphical user interface consists of four parts: protein table (top left), peptide table (bottom left), graph pane (top right), and text pane (bottom right). The graph pane contains seven graphs: Likelihood Plot, Sequence Coverage Plot, Ion Chromatogram, PPC Chromatogram, PCA Plot, MS1 Scan, and MS2 Scan. All graphs can be detached from the graph pane.



Protein Quantification Results

	Locus	Log2R	CI	Description
1	RPA2164	-0.1	0.3	groEL2 chaperonin GroEL2, cpn60 2437219:2438862 reverse MW 57798
2	RPA3268	-0.2	0.4	rpoB RNA polymerase beta subunit 3683017:3687141 reverse MW 153601
3	RPA2047	-0.1	0.4	dctP1 TrapT family, dctP subunit, C4-dicarboxylate periplasmic binding prote...
4	RPA1705	-0.3	0.4	RPA1705 possible 3-Hydroxyacyl-CoA dehydrogenase type II 1903489:19045...
5	RPA3297	0.1	0.4	RPA3297 possible branched-chain amino acid transport system substrate-bin...
6	RPA0333	-0.2	0.4	dnaK heat shock protein DnaK (70) 360754:362649 forward MW 67947
7	RPA2153	-0.5		50503
8	RPA3724	-0.2		eriplasmi...
9	RPA3253	-0.2		
10	RPA3267	-0.3		156056
11	RPA0432	-0.1		e MW 7...
12	RPA4308	-0.2		reverse ...
13	RPA0064	-0.3		
14	RPA4216	-0.2		
15	RPA0192	0		
16	RPA4034	0.1		nding pr...
17	RPA0531	-0.2		72 forwar...
18	RPA1140	-0.2		57626
19	RPA0944	-0.1		8:10441...
20	RPA0943	0		4
21	RPA4721	0		423:532...
22	RPA0176	-0.3		204 rev...
23	RPA4643	-0.2		
24	RPA2667	-0.2		W 26287
25	RPA2874	-0.2		
26	RPA0191	0		W 42319
27	RPA2921	-0.2		
28	RPA4050	-0.2		
29	RPA2200	-0.2	0.6	guaB inosine monophosphate dehydrogenase 2475579:2477074 forward M...
30	RPA1789	-0.3	0.6	RPA1789 putative branched-chain amino acid transport system substrate-bin...
31	RPA1535	0	0.6	cycA cytochrome c2 1703088:1703507 reverse MW 14633
32	RPA4749	-0.4	0.6	etfA electron transfer flavoprotein alpha-subunit, (ETFLS) 5360445:5361389 r...

Configuration - ProRata

Denominator Isotope | Numerator Isotope | Quantification | Chromatogram

Residue Atomic Composition Table

Name: reference   Add PTM   Insert Row

	C12	H1	O16	N14	P31	S32	C13	H2	O18	N15
NTerm	0	1	0	0	0	0	0	0	0	0
CTerm	0	1	1	0	0	0	0	0	0	0
L	6	11	1	1	0	0	0	0	0	0
A	3	5	1	1	0	0	0	0	0	0
S	3	5	2	1	0	0	0	0	0	0
G	2	3	1	1	0	0	0	0	0	0
V	5	9	1	1	0	0	0	0	0	0
E	5	7	3	1	0	0	0	0	0	0
K	6	12	1	2	0	0	0	0	0	0
I	6	11	1	1	0	0	0	0	0	0

Cancel   Save...

Peptide Quantification Results

	Sequence	Log2R	log2SN	Validity
427	R.GVDILANAVK.V	-0.170369	3.25825	true
428	K.LAGGVAVIR.V	-0.0689786	3.33083	false
429	K.VAAVKAPGFG...	-0.350363	3.35168	true
430	K.MENVTLQMLG...	-0.0126063	3.49402	true
431	K.MENVTLQMLG...	-0.0946022	3.58811	true
432	K.AMLQDIAILTG...	0.62227	3.61348	true
433	K.EQYAFGFDSQ...	-7.28808	3.64597	true
434	R.VEFDDAYILIN...	-0.0446026	3.70518	true
435	K.KGIIDPTK.V	-0.0258911	3.9285	true
436	R.TAIQNAASVAA...	-3.24845	4.06286	true
437	K.SLETELDVVEG...	4.95854	4.06761	true
438	R.RALSAPAR.Q	0.0232154	4.21671	true
439	K.KGIIDPTK.V	-0.194036	4.21853	true

Protein Information

Protein description groEL2 chaperonin GroEL2, cpn60 2437219:2438862 reverse M

Abundance ratio: treatment : reference

Abundance ratio: 0.933033 Log2(abundance ratio) -0.1 Confidence interval [-0.3, 0]

Quantification hits 186   Identification hits 439 Locus RPA216

Peptide Information

Sequence K.KGIIDPTK.V

Identification filename N14_031405_RPal_N14N15_1to1_LTQ_07.2063.2063.1
N14_031405_RPal_N14N15_1to1_LTQ_07.2075.2075.1
N15_031405_RPal_N14N15_1to1_LTQ_07.2050.2050.1
N15_031405_RPal_N14N15_1to1_LTQ_07.2064.2064.1

Chromatogram filename 031405_RPal_N14N15_1to1_LTQ_07.1.100.xic.xml

XIC identifier 88   Peak shift 0   ID score

PCA abundance ratio 0.874157   Peak area ratio 0.879638   Peak he

Profile S/N 18.6167   reference S/N 228.462   treatmen

Full Scan Range [1899, 2247] treatment m/z range [871.017, 874.023] referenc

FDR could exceed 10% with one replicate, but was less than 5% or 3% with two or three replicates, respectively. The t-test performed the best with three replicates. Third, we generalized the G test to increase the sensitivity of detecting differential protein expression under multiple experimental conditions. Out of 1,664 detected *R. palustris* proteins in the LC-MS/MS experiment, the generalized G test identified 1,119 differentially expressed proteins under six growth conditions, including photoheterotrophic, chemoheterotrophic, nitrogen fixation, photoautotrophic, stationary phase, as well as benzoate as an alternate carbon source. Unlike 2-fold change under two conditions, the generalized G test differentiated 300 more proteins under six conditions. Furthermore, among the 625,521 protein pairs between these 1,119 differentially expressed proteins, operon pairs were much stronger co-expressed than the non-operon ones. Finally, we identified six protein clusters with known biological significance by combining cluster analysis with functional annotation of these differentially expressed proteins. In summary, the proposed generalized G test using spectral count sampling statistics is a viable methodology for robust quantification of relative protein abundance and for sensitive detection of biologically significant differential and correlated protein expression under multiple experimental conditions in label-free shotgun proteomics.

Research sponsored by the DOE Genomics:GTL Program.

56

High Throughput Comprehensive and Quantitative Microbial Proteomics: Production in Practice

Richard D. Smith* (rds@pnl.gov), Joshua N. Adkins, David J. Anderson, Gordon A. Anderson, Kenneth J. Auberry, Michael A. Buschbach, Stephen J. Callister, Therese R.W. Clauss, Jim K. Fredrickson, Kim K. Hixson, Navdeep Jaitly, Gary R. Kiebel, Mary S. Lipton, Eric A. Livesay, Matthew E. Monroe, Ronald J. Moore, Heather M. Mottaz, Angela D. Norbeck, Daniel J. Orton, Ljiljana Paša-Tolić, Kostantinos Petritis, David C. Prior, Samuel O. Purvine, Yufeng Shen, Anil K. Shukla, Aleksey V. Tolmachev, Nikola Tolić, Harold R. Udseth, Rui Zhang, and Rui Zhao

Pacific Northwest National Laboratory, Richland, WA

Significance: Capabilities for quantitative proteomics measurements of steadily increasing throughput and quality have been implemented and are being applied to studies with a range of microbial systems.

With recent advances in whole genome sequencing for a growing number of organisms, biological research is increasingly incorporating higher-level "systems" perspectives and approaches. Key to supporting systems-level advances in microbial and other biological research at the heart of the DOE Genome GTL program is the ability to quantitatively measure the array of proteins (i.e., the proteome) for various organisms under many different conditions.

Among the challenges associated with making useful comprehensive proteomic measurements are identifying and quantifying large sets of proteins whose relative abundances span many orders of magnitude. Additionally, these proteins may vary broadly in chemical and physical properties, have transient and low levels of modifications, and be subject to endogenous proteolytic processing. Ultimately, such measurements and the resulting insight into biochemical processes are expected to enable development of predictive computational models that could profoundly affect environmental clean-up, understandings related to climate, and energy production by e.g., providing a more solid

* Presenting author

basis for mitigating the impacts of energy production-related activities on the environment and human health.

A “prototype high throughput production” lab established in FY 2002 was an early step towards implementing higher throughput proteomics measurements. Operations within this lab remain distinct from technology development efforts, both in laboratory space and staffing. This step was instituted in recognition of the different staff “mind sets” required for success in these different areas, as well as to allow “periodic upgrades” of the technology platform in a manner that does not significantly impact its production operation. The result has been faster implementation of technology advances and more robust automation of technologies that improve overall effectiveness.

The biological applications of the technology and associated activities are the subject of a separate, but interrelated project (J. K. Fredrickson, PI), involving studies of a number of microbial systems (e.g., *Shewanella oneidensis*, *Geobacter sulfurreducens, Rhodobacter sphaeroides*) in collaboration with leading experts on each organism. These studies have demonstrated the capability for automated high-confidence protein identifications, broad proteome coverage, and for exploiting both stable isotope labeling and label-free methods to obtain high precision in protein abundance measurements.

With a paradigm established for high throughput proteomic measurements, our primary goal now is to significantly increase data quality, as well as throughput. A significant challenge is how to maximize the information content derived from large and complex data sets such that the researcher can gain novel biological insights. Thus, a key component of our program involves developing the informatics tools needed to quantify and define the quality of data, as well as the tools to make the results broadly available and understandable to the researchers. Efforts currently in progress aim to:

- Significantly increase the overall data production by more than an order of magnitude in conjunction with increased data quality, providing data that are quantitative and have statistically-based measures of quality.
- Extend the application to an increasing number of different kinds of post-translation modifications.
- Provide the infrastructure and informatics tools required to efficiently manage, use, and disseminate large quantities of data generated by GTL “users.”

This presentation will highlight the advances in providing high quality data with statistically-founded measures of quality, while providing increased measurement throughput. The advances will be illustrated in the context of applications to microbes of interest to the GTL program.

Acknowledgements: This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute through Contract No. DEAC05-76RLO 1830.

57

H_2O_2-Induced Stress Responses of *Shewanella oneidensis* MR-1

T. Li[1,2], J. Guo[2], D. Klingeman[1], L. Wu[2], X. Liu[3], T. Yan[2], Y. Xu[2], A. Beliaev[4], Z. He[5,7], T.C. Hazen[6,7], **A. P. Arkin**[6,7], J. Zhou[1,5,7]* (jzhou@ou.edu)

[1]Oak Ridge National Laboratory, Oak Ridge, TN; [2]University of Georgia, Athens, GA; [3]Central South University, Changsha, China; [4]Pacific Northwest National Laboratory, Richland, WA; [5]University of Oklahoma, Norman, OK; [6]Lawrence Berkeley National Laboratory, Berkeley, CA; and [7]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

Shewanella oneidensis is a facultative anaerobe that possesses a complex electron-transport system, which allows the coupling of metal reduction to bacterial energy generation. The ability of *S. oneidensis* to reduce toxic heavy metals makes it an ideal candidate for bioremediation of contaminated sites. However, many questions remain about how the organism responds to and functions in various environmental conditions. To better understand the molecular biology of this metal-reducing bacterium, we have employed whole-genome microarrays and bioinformatics to investigate the global gene expression profiles of wild type and oxyR mutant strains in response to H_2O_2 induced oxidative stress. In wild type *Shewanella* cells, a total of 1,092 genes showed significant changes in the expression level under at least one condition tested. Many genes showed dose-dependent expression pattern and were differentially regulated at different time points. Comparison of the gene expression kinetics suggests that *S. oneidensis* possesses complex regulatory systems to protect the cells from oxidative stress. Among the genes that are immediately up regulated under stress conditions are the genes with established functions in oxidative stress, such as the alkyl hydroperoxide reductase (Ahp) gene, the catalase (Kat) genes, and the stress response DNA-binding protein (dps) gene, as well as various genes that have not been previously described to be involved in the oxidative stress responses of other bacterial species, including some iron- and sulfur- responsive genes. In addition, an oxyR homologue is identified and characterized. Phenotypic and enzymatic studies along with microarray analysis indicate that *S.*oneidensis oxyR serves as a dual function transcription regulator, which activates the expression of many oxidative stress genes in response to H_2O_2 stress while repressing the expression of a catalase gene (KatB) and the nonspecific DNA binding protein gene (dps) under normal growth condition.

Iron is an essential nutrient with limited bioavailability. Due to its ability to react with and catalyze the generation of toxic radicals, iron when overloaded may pose a big threat to living cells and tissues; and so the acquisition of iron is usually tightly regulated in biological systems. In contrast to *E. coli*, the wild type *Shewanella oneidensis* fur gene is not significantly affected by H_2O_2, while many iron inducible genes showed drastic up-regulation under H_2O_2 stress conditions. These genes include the TonB1 iron transport system genes, a bacterioferritin gene bfr, the ferrous iron transport genes, and some DNA repair and metabolism genes. To investigate the H_2O_2-induced regulation of the iron responsive genes in *S. oneidensis*, we also included a fur deletion mutant in our microarray studies. Not surprisingly the fur deletion mutant showed hypersensitivity to H_2O_2 stress. Further analysis by microarray study of the fur mutant under oxidative stress condition reveals interesting regulation pattern. In fur mutant the iron responsive genes are highly expressed when compared with the wild type strain, which is in agreement with the repressor role of fur. However in contrast to wild type strain fur mutant showed little or diminished induction of the iron inducible genes in response to H_2O_2, indicating that fur directly or indirectly mediates the upregulation of these genes by H_2O_2 treatment, and the regulation is apparently independent of the transcription of the fur gene since the expression level of fur does not show significant change under the stress condition. More interestingly, many of the fur regulated iron responsive genes also exhibit unusually high expression in oxyR deletion

* Presenting author

mutant, and when H_2O_2 is applied to the oxyR mutant the transcription of the iron genes further increases, which coincides well with the regulation pattern of these genes in wild type strain; these data suggest that oxyR mutation may play a role in modulating fur activity. At the time being we are applying bioinformatics.

58

Application of High Resolution Proteomics to Characterize Microbial Systems for Metal Reduction and Photosynthesis

Mary S. Lipton[1]* (mary.lipton@pnl.gov), Margie F. Romine[1], Stephen J. Callister[1], Kim K. Hixson[1], Samuel O. Purvine[1], Angela D. Norbeck[1], Joshua N. Adkins[1], Matthew E. Monroe[1], Yuri A. Gorby[1], Carrie D. Goddard[1], Richard D. Smith[1], **Jim K. Fredrickson**[1], Derek Lovley[2], Richard Ding[2], Timothy Donohue[3], Miguel Dominguez[3], Christine Tavano[3], Samuel Kaplan[4], Xiaihua Zeng[4], Jung Hyeob Roh[4], and Frank Larimer[5]

[1]Pacific Northwest National Laboratory, Richland, WA; [2]University of Massachusetts, Amherst, MA; [3]University of Wisconsin, Madison, WI; [4]University of Texas Medical School, Houston, TX; and [5]Oak Ridge National Laboratory, Oak Ridge, TN

Exploiting microbial function for purposes of bioremediation, energy production, carbon sequestration and other missions important to the DOE requires an in-depth and systems level understanding of the molecular components of the cell that confer its function. Inherent to developing this improved understanding is the ability to rapidly acquire global quantitative measurements of the proteome (i.e. the proteins expressed in the cell). We have applied our state-of-art proteomics technologies based on high-resolution separations combined with Fourier transform ion cyclotron resonance mass spectrometry to obtain quantitative and high throughput global proteomic measurements of *Shewanella oneidensis*, *Geobacter sulfurreducens* and *Rhodobacter sphaeroides.*

Accurate ORF identification and functional annotation are important for post-genome analysis of any organism. The *Shewanella* Federation has organized an effort to update and refine the predictions of RNA and protein-encoding genes (CDS) for *Shewanella oneidensis* MR-1. The proteome data has proven to be an exceptional resource for validating hypothetical protein predictions, as well as for positioning start codons and signal peptidase cleavage sites. 1197 hypothetical proteins were manually re-evaluated to access the robustness of predictions. As a result of this evaluation, 525 CDS predictions were dropped and 225 were validated. We have also initiated mining of the *Shewanella* proteome data to validate start codons. Initial analyses in which we used the original predicted start codons validated 801 N-termini. Our results suggest 306 of these termini are a consequence of cleavage of the N-terminal amino acid by methionylaminopeptidase and that the predominant penultimate amino acids found corresponded well with those found in *E. coli*. Analyses are currently underway to mine proteome data for evidence of 1) pseudogene expression, 2) additional start codon validation based on new start predictions, 3) leader peptidase cleavage sites, and 4) identification of new CDS.

Geobacter sulfurreducens is a representative of an important genus of metal-reducing bacteria that predominate in a variety of subsurface environments in which Fe(III) oxide reduction is important. Changes in protein expression levels were investigated in *Geobacter* cultured with different terminal electron acceptors. The abundance of proteins in various subcellular fractions of *Geobacter sulfurreducens* grown on fumarate or Fe(III) citrate were determined and the results compared to identify proteins associated with these distinct modes of anaerobic respiration. Among the proteins that

* Presenting author

changed, 91 *c*-type cytochromes were identified. Relative abundance of some *c*-type cytochromes varied markedly with different growth conditions. Higher abundance of cytochromes during growth on Fe(III) may be indicative of cytochromes that play an essential role in Fe(III) reduction. To better understand the physiology of *Geobacter* species during growth on Fe(III) oxide, the proteome of *G. sulfurreducens* grown on Fe(III) oxide was compared with the proteome of cells grown with soluble Fe(III) citrate. Analysis using the accurate mass and time tag (AMT) approach revealed many *c*-type cytochromes that were significantly more abundant in cells grown with insoluble Fe(III) oxide when compared with cells grown on soluble Fe(III) citrate as the electron acceptor. These cytochromes included the outer-membrane *c*-type cytochrome, OmcS and OmcG, all of which genetic studies have suggested are required for Fe(III) oxide reduction. Furthermore, several other uncharacterized cytochromes were determined to be significantly up regulated during growth on Fe(III) oxide. A number of other proteins of unknown function were also more abundant during growth on Fe(III) oxide than on soluble Fe(III).

Rhodobacter sphaeroides 2.4.1 is capable of growth under a variety of conditions and has been studied in relation to its ability to reduce metals, fix nitrogen, sequester carbon dioxide, and produce energy through photosynthesis. The complex sensory and regulatory network responsible for the transition of *R. sphaeroides* to a different metabolic steady state suggests the presence of proteins directly and indirectly involved in the photosynthetic lifestyle; beyond the structural and regulatory proteins transcribed and translated from the well-characterized photosynthetic gene cluster (PGC). We present results that characterize the proteome of aerobic and photosynthetic cell cultures by utilizing: 1) proteins extracted from whole cell lysate, soluble, insoluble, and global fractions, and 2) proteins extracted from sub-cellular fractions that include cytoplasm, cytoplasmic membrane, periplasm, outer membrane, and chromatophore. Both analyses utilized the AMT approach. The first analysis emphasized the role of observed proteins in the photosynthetic lifestyle of *R. sphaeroides*, such as those involved in electron transport, and compared results with available transcriptome data. The second analysis emphasized the localization of proteins within the cellular matrix. Localization for many of the 4269 proteins predicted from *R. sphaeroides'* sequenced genome has not been characterized beyond that implied by prediction algorithms and functional annotation. Therefore, it is important to determine the localization of the proteins within the organism to achieve a clear view of the physiology.

Acknowledgements: This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute through Contract No. DEAC05-76RLO 1830.

* Presenting author

59

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

Terry C. Hazen[1,8]* (tchazen@lbl.gov), Carl Abulencia[3,8], Gary Andersen[1], Sharon Borglin[1,8], Eoin Brodie[1], Steve van Dien[7], Matthew Fields[6,8], Jil Geller[1,8], Hoi-Ying Holman[1,8], Rick Huang[1,8], Richard Phan[1], Eleanor Wozei[1], Janet Jacobsen[1,8], Dominique Joyner[1,8], Romy Chakraborty[1,8], Martin Keller[3,8], Aindrila Mukhopadhyay[1,8], David Stahl[5,8], Sergey Stolyar[5,8], Judy Wall[4,8], Denise Wyborski[3,8], Huei-che Yen[4,8], Grant Zane[4,8], Jizhong Zhou[2,8], E. Hendrickson[5], T. Lie[5], J. Leigh[5], and Chris Walker[5,8]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]Oak Ridge National Laboratory, Oak Ridge, TN; [3]Diversa, Inc., San Diego, CA; [4]University of Missouri, Columbia, MO; [5]University of Washington, Seattle, WA; [6]Miami University, Oxford, OH; [7]Genomatica, Inc., San Diego, CA; and [8]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

Field Studies

Environmental Characterizations. Nitrate and heavy metals (including uranium and technetium) are a major groundwater contaminant at the U.S. Department of Energy (DOE) NABIR-Field Research Center in Oak Ridge, TN. The sites are marked by acidic conditions, high concentrations of nitrate, chlorinated solvents, and heavy metals. Groundwater bacterial communities were monitored in several wells along a transect that were stimulated via the addition of a potential electron donor (i.e., ethanol). Electron donor was added intermittently over 650 days. By day 535, the nitrate levels in the groundwater had decreased from 10 mM to 0.5 mM, and groundwater uranium levels had declined from approximately 2 mg/l to 0.2 mg/l. Bacterial community composition and structure were characterized via clonal libraries of the SSU rRNA gene sequences. The up-stream and injection well had similar diversity indices, whereas the treatment zone and immediately down-stream well both had increased diversity. When the entire sequence libraries were compared via LIBSHUFF analysis. The results indicated that the bacterial community composition and structure changed upon bio-stimulation for metal-reducing conditions, and that sequences indicative of *Anaeromyxobacter* and *Desulfovibrio* were detected in wells that displayed a decline in both nitrate and uranium upon bio-stimulation. The results also suggested that, in addition to the presence of desired populations, an increase in diversity may be important for optimal functionality.

^{13}C-labelled lactate was injected in August 2004 at the Hanford 100H site to biostimulate chromium reduction. After more than 1 year chromium was still at non-detect in the stimulated wells. 16s phylochip analyses showed a dramatic increase in diversity at the stimulated wells, including iron reducers (*Geobacter*) and sulfate reducers (*Desulfovibrio*). Sequentially competing terminal electron acceptors were depleted: oxygen, nitrate, iron(III), and sulfate. Methane however was never detected, though ^{13}C was detected in the dissolved inorganic carbon and in the signature lipids (PLFA) of iron reducers and sulfate reducers. Sulfate reduction was still active after more then a year in the deepest parts of the aquifer, and iron(II) still dominated suggesting an active Cr(IV) reducing environment. *Desulfovibrio* strains have been isolated and are currently being sequenced. Stress responses in these strains will be compared to the pipeline studies on DvH already completed.

Biopanning/Clone libraries. Low biomass samples from nitrate and heavy metal contaminated soils yield DNA amounts which have limited use for direct, native analysis and screening. Multiple displacement amplification (MDA) using ϕ 29 DNA polymerase was used to amplify whole genomes from environmental, contaminated, subsurface sediments. By first amplifying the gDNA, biodiversity

* Presenting author

analysis and genomic DNA library construction of microbes found in contaminated soils were made possible. Whole genome amplification of metagenomic DNA from very minute microbial sources enables access to genomic information that was not previously accessible. DNA was extracted from nine subsurface soil samples from five different areas within the DOE NABIR FRC site. The samples represent different geographical areas containing various levels of contaminants and varying subsurface depths. Multiple displacement amplification was used to amplify whole genomes from the extracted DNA. By first amplifying the gDNA, biodiversity analysis and genomic DNA library construction of microbes found in these contaminated soils were made possible. After amplification, SSU rRNA analysis revealed relatively even distribution of species across several major phyla. Clone libraries were constructed from the amplified gDNA and a small subset of clones was used for shot gun sequencing. BLAST analysis of the library clone sequences, and COG analysis, showed that the libraries were diverse and the majority of sequences had sequence identity to known proteins. The libraries were screened by DNA hybridization and sequence analysis for native histidine kinase genes. 37 clones were discovered that contained partial histidine kinase genes, and also partial, associated response regulators and flanking genes.

Enrichments. This year we also isolated and characterized a new strain of *D. vulgaris* that should be valuable as a genetic tool and to investigate the role of bacteriophage in microbial stress responses. *D. vulgaris* DePue was isolated from sediments of the metal (Zn) contaminated Lake DePue. Our analysis revealed that this strain did not possess phage genes that were found in the *D. vulgaris* Hildenborough genome and was susceptible to two phages from the former strain. Currently, genome sequencing of strain DePue is underway at the JGI.

Dual culture systems. Although sulfate-reducing bacteria (SRB) characteristically respire sulfate, their distribution does not appear to be restricted by sulfate availability. In the absence of sulfate, some SRBs can grow by cooperating syntrophically with hydrogenotrophic methanogens. We established and characterized a syntrophic coculture between the model SRP *Desulfovibrio vulgaris* Hildenborough and *Methanococcus maripaludis* in order to study the physiology and influence of stressors on the growth of *D. vulgaris* in the absence of sulfate. In this interaction, the species must cooperate to transform lactate and carbon dioxide into acetate and methane by transferring reducing equivalents. Interspecies hydrogen transfer is known as a driving process for such interactions, but transfer of formate or other compounds might also occur. Using publicly available genomic and physiological data, we developed a stoichiometric metabolic model that predicts that both hydrogen and formate could be used as electron carriers. This model was able to predict metabolite accumulation of *D. vulgaris* in monoculture and in the syntrophic coculture at multiple stages in their dynamic growth cycle, but there were discrepancies between the experimental data and model predictions at individual time points. These discrepancies may be explained by dynamic oscillations of growth of each species in coculture, a phenomenon which is supported by a dramatically changing ratio of 16S rRNA production of each species over time. We have also compared gene expression of each species under coculture and monoculture conditions. Although experiments with mutants show that it is possible for coculture growth to occur when the capacity to transfer electrons via formate has been deleted, initial gene expression data and the metabolic model both suggest that formate may be used as an electron carrier under syntrophic conditions. Further comparisons of mutant cocultures will be performed to address this issue. Gene expression analyses of co-cultures has revealed several interesting gene expression changes in *D. vulgaris.* Even though there was no sulfate available, *D. vulgaris* genes for the sulfate reduction pathway were expressed. Genes for conversion of lactate to acetate and acetate excretion were downregulated during syntrophic growth compared to mono-culture growing in the presence of sulfate even though lactate was the sole carbon source in each condition. Finally, differential expression of several hydrogenases thought to be important for *D. vulgaris* growth under the different conditions were detected.

Stress Experiments

High Throughput Biomass Production. Producing large quantities of high quality and defensibly reproducible cells that have been exposed to specific environmental stressors is critical to high throughput and concomitant analyses using transcriptomics, proteomics, metabolomics, and lipidomics. Culture of *D. vulgaris* is made even more difficult because it is an obligate anaerobe and sulfate reducer. For the past three years, our Genomics:GTL VIMSS project has developed defined media, stock culture handling, scale-up protocols, bioreactors, and cell harvesting protocols to maximize throughput for simultaneous sampling for lipidomics, transcriptomics, proteomics, and metabolomics. All cells for every experiment, for every analysis are within two subcultures of the original ATCC culture of *D. vulgaris*. In the past three years we have produced biomass for 80 (40 in the last year) integrated experiments (oxygen, NaCl, NO_3, NO_2, heat shock, cold shock, pH, Cr, and mutants Fur and Per) each with as much as 30 liters of mid-log phase cells (3 x 10^8 cells/ml). In addition, more than 60 adhoc experiments for supportive studies have been done each with 1-6 liters of culture. All cultures, all media components, all protocols, all analyses, all instruments, and all shipping records are completely documented using QA/QC level 1 for every experiment and made available to all investigators on the VIMSS Biofiles database (http://vimss.lbl.gov/perl/biofiles). To determine the optimal growth conditions and determine the minimum inhibitory concentration (MIC) of different stressors we adapted plate reader technology using Biolog and Omnilog readers using anaerobic bags and sealed plates. Since each well of the 96-well plate produces an automated growth curve, over more than 200 h, this has enabled us to do more than 6,000 growth curves over the last two years. Since the Omnilog can monitor 50 plates at a time, this allows us to do more than 5,000 growth curves in a year.

Phenotypic Responses. Phenotypic Microarray™ analysis is a recently developed analytical tool to determine the phenotype of an organism. In the last year we have further refined our phenotyping of DvH to minimize the number of plates necessary. We have also screened 15 knockout mutants of DvH and 10 knockout mutants of *Shewanella* MR1. See (https://vimss.lbl.gov/~jsjacobsen/cgi-bin/Test/HazenLab/Omnilog/home.cgi) for sample data sets and analyses.

Synchrotron FTIR Spectromicroscopy for Real-Time Stress Analysis. This year the stress responses in *Desulfovibrio vulgaris* triggered by oxygen (O_2), nitrate (NO_3), Cr, and sodium chloride (NaCl) were studied using FTIR. The advantage of the FTIR spectroscopy approach is that it allows us to immediately detect *in situ* intracellular molecules or molecular structures, to nondestructively monitor and quantify metabolites produced in response to different stresses, and to rapidly characterize growth-dependence phenomena and stress-response mechanisms. Because the chemical and structural information of molecules associated with cellular processes inside microbes are contained in each infrared spectrum, one can extract chemical and structural information from each spectrum regarding the physiological conditions of a cell or a group of cells. By comparing measurements, we were able to identify tight temporal changes in chemical bonds, functional groups, and chemical substructures in lipids, DNA, proteins, and polyglucose in *D. vulgari*s. For example, when exposed to moderate concentrations of O_2 or NO_3, *D. vulgaris* increases the production of exopolysaccharides but with little change in protein structures. However, when exposed to moderate concentration of NaCl, *D. vulgaris* again increases the production of exopolysaccharides while exhibiting a significant change in protein structures. These results, together with microscopy images, confirmed the importance of exopolysaccharide production in enhancing the stress resistance and survival of *D. vulgaris*. These studies also enabled focusing of VIMSS transcriptomic, proteomic, and metabolomic studies on the best time points to rapidly resolve stress response pathways.

60

The Virtual Institute of Microbial Stress and Survival (VIMSS): Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Microbes

Carl Abulencia[4,8], Eric Alm[1,8], Gary Andersen[1], **Adam P. Arkin**[1,8]* (APArkin@lbl.gov), Kelly Bender[5,8], Sharon Borglin[1,8], Eoin Brodie[1], Romy Chakraborty[1,8], Swapnil Chhabra[3,8], Steve van Dien[6], Inna Dubchak[1,8], Matthew Fields[7,8], Sara Gaucher[3,8], Jil Geller[1,8], Masood Hadi[3,8], Terry W. Hazen[1,8], Qiang He[2], Zhili He[2,8], Hoi-Ying Holman[1,8], Katherine Huang[1,8], Rick Huang[1,8], Janet Jacobsen[1,8], Dominique Joyner[1,8], Jay Keasling[1,8], Keith Keller[1,8], Martin Keller[4,8], Aindrila Mukhopadhyay[1,8], Richard Phan[1,8], Morgan Price[1,8], Joseph A. Ringbauer Jr.[5,8], Anup Singh[3,8], David Stahl[6,8], Sergey Stolyar[6,8], Jun Sun[4], Dorothea Thompson[2,8], Christopher Walker[6,8], Judy Wall[5,8], Jing Wei[4], Denise Wolf[1,8], Denise Wyborski[4,8], Huei-che Yen[5,8], Grant Zane[5,8], Jizhong Zhou[2,8], and Beto Zuniga[6]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]Oak Ridge National Laboratory, Oak Ridge, TN; [3]Sandia National Laboratories, Livermore, CA; [4]Diversa, Inc., San Diego, CA; [5]University of Missouri, Columbia, MO; [6]University of Washington, Seattle, WA; [7]Miami University, Oxford, OH; and [8]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

Introduction

The mission of the Virtual Institute of Microbial Stress and Survival is to understand the molecular basis for the survival and growth of microbes in the environment. Towards this end VIMSS has designed a series of key protocols, experimental pipelines and computational analyses to support and coordinate research in this area. Our flagship project aims to elucidate the pathways and community interactions which underlie the ability of *Desulfovibrio vulgaris* Hildenborough (DvH) to survive in diverse, possibly contaminated environments and reduce metals. Their ability to reduce toxic Uranium and Chromium, major contaminants of industrial and DOE waste sites, to a less soluble form has made them attractive from the perspective of bioremediation.

We are discovering the molecular basis for the physiology of these organisms first through characterization of the biogeochemical environment in which these microbes live and how different features of these environments affect their growth and reductive potential. We have created an integrated program through the creation of an experimental pipeline for the physiological and functional genomic characterization of microbes under diverse perturbations. This pipeline produced controlled biomass for a plethora of analyses as described below and is managed through workflow tools and a data management and analysis system. The effort is broken into three interacting core activities: The Applied Environmental Microbiology Core; the Functional Genomics Core; and the Computational Core. While the individual accomplishments of these cores may be found in more detail, we summarize the highlights here.

Accomplishments of the Applied Environmental Microbiology Core (AEMC)

Characterization of the Environment. The AEMC has been monitoring natural and stimulated groundwater and soil communities in DOE NABIR Field Research sites. These sites are contaminated with different combinations of heavy metals, low pH, chlorinated solvents, nitrates and other cellular stressors. Strong correlations between the population growth of certain strains of bacteria including DvH like strains and metal reduction dynamics were observed. *D. vulgaris* strains from multiple sites have been enriched and isolated and some are undergoing sequencing now. A novel cloning method was used to amplify DNA and whole genomes from five separate contaminated sites

* Presenting author

to compare and contrast functional diversity among these sites and relate this diversity to the stability and metal-reduction efficacy of the communities at these sites (including *D. vulgaris).* An initial survey of the signal transduction genes is underway.

Biomass production and Characterization: In the core pipeline experiments each microbe is first characterized physiologically using Omnilog phenotypic microarrays. A stressor condition is then applied to a large set of batch cultures and samples are collected periodically to obtain a time-series of cellular response. Each time-point is split so that the cells can be imaged, analyzed through synchrotron IR microscopy to measure the bulk physiological changes of the cells during their response, and determine the optimal time points to send to the functional genomics core (FGC) for transcript, protein and metabolite analysis. Twenty-three conditions in DvH and *Shewanella oneidensis* (So) have been produced including mutant studies and co-culture of DvH syntrophically coupled to a hydrogenotrophic methanogen (*Methanococcus maripaludis*). Using publicly available genomic and physiological data, we developed a stoichiometric metabolic model that predicts that both hydrogen and formate could be used as electron carriers. This model was able to predict metabolite accumulation of *D. vulgaris* in monoculture and in the syntrophic coculture at multiple stages in their dynamic growth cycle, but there were discrepancies between the experimental data and model predictions at individual time points which can be explained by interesting observed growth dynamics.

Accomplishments of the Functional Genomics Core

Genetics: Our bar-coded deletion project proceeds apace and a number of deletions have been phenotyped by omnilog array and through the VIMSS physiological pipeline. New affinity tags have been developed for using in pull-down and molecular complex studies both for this and a collaboratoring project. We have also expanded our transposon mutagenesis library and have begun to array them for sequencing and phenotyping. Ultimately the most interesting of these will be submitted to pipeline studies.

Transcriptomics: We have, to date, characterized seventeen stresses, growth phase conditions, or mutant responses in *DvH* and six in *S. oneidensis* and results are integrated with the VIMSS MicrobesOnline Database. New regulons and their cis-regulatory sequences have been discovered along with new hypotheses of the pathways by which both organisms respond to these different stressors. A number of papers are in press, submitted or are in preparation around this topic. Compendium analysis for DvH is underway. *Geobacter* stressors are coming online and will be used for the three organism cross comparison.

Proteomics. As part of the proteomics mission we have developed and compared and contrasted MS/MS, ICAT MS, ITRAQ_MS, and DIGE (Differential In-gel Electrophoresis) MALDI to profile the protein abundance and protein abundance changes in response to stressors. The year has been spent in quantifying reproducibility and accuracy of these results and comparing their predictions to that of microarray. There were significant differences among the methods and with the microarray data, however, these highlight the different sensitivities of the instruments and the long cell-cycle times of the microorganism. Recent, longer term experiments show better agreement between the microarray and proteomic data. We are now targeting our proteomics efforts to track the changes in the key sulfate reducer signature genes (see below) and their linked stress response pathway regulators to generate a more limited but precise measure of both abundance and redox changes.

Metabolomics: We have set up and optimized both Capillary electrophoresis (CE) and Liquid chromatography (LC) coupled with Mass spectrometry (MS) methods for characterization of metabolites. Metabolite extraction protocols have been developed for *DvH.* A new Fourier transform ion cyclotron resonance mass spectrometer has recently come online which will allow a much wider survey of metabolites without the need for external standards. However, we are still in testing phase of this technology. We have identified the key pathways and metabolites on which we will focus to understand the sulfate-reduction and stress metabolism of *DvH* in contrast with *S. oneidensis* and *Geobacter.*

Accomplishments of the Computational Core

The computational core has continued its core development of MicrobesOnline (web site: http://MicrobesOnline.org) as the core framework for comparative functional microbial genomics and for the dissemination and visualization of VIMSS data. The microarray database is now fully integrated with the website and proteomic, metabolomic and molecular interaction data will shortly follow. A highly curated database of cis-regulatory motifs and their regulators in a variety of organisms from Mikhail Gelfand is undergoing integration with MicrobesOnline and a suite of cis-regulatory prediction tools are being added to the informatics tools available on the site. The team has used the site to provide nearly automated data analysis for the microarray experiments above and has been used to understand the key biological mechanisms in a number of responses including cold and heat shock, low and high pH, low and high oxygen, salt and osmotic shock and adaptation, nitrate and nitrite stress, chromium reducing conditions, growth in co-culture with a methanogen, iron limitation and mutation in the fur gene. A new method called OpWise has aided in this analysis and uses prior computational core work in accurate operon prediction to exploit genomic architecture to get better estimates of gene expression changes and the systematic and non-systematic errors underlying the measurements. We are now in the process of integrating the new proteomic and metabolomic data into the framework. We already have a new interface for analysis and visualization for the phenotype microarray data which we are now using for quality control and analysis of mutant data.

Along with the core platform development and data analysis and interpretation, the computational core has used MicrobesOnline as a platform for more general discovery and annotation. MicrobesOnline is being used as an annotation tool for a number of new genomes and environmental sequences. For example, in collaboration with another project we have identified and annotated a novel sulfate reducing organisms isolated from a deep South African Goldmine. In addition, a set of signature genes that define sulfate reducing bacteria and Archea were inferred and validated against the gene expression database. A set of *DVH* centered comparative pathway analyses for the Metabolic, global regulatory, dissimilatory nitrogen oxide pathways were accomplished with Mikhail Gelfand's group. New theories for the formation, maintenance, tuning, and death of operon structure were put forward and a comprehensive study on the evolution of histidine kinases was completed. This latter study also discovered that certain organisms are more likely to generate new kinases by lineage specific expansion and other to acquire them through horizontal gene transfer.

Future Work

In this next year we are focusing on integrating our stress condition measurements and models to the ability of the organism to reduce metals. We will relate the resultant model to the ability of the organism to survive and reduce metals in the various environments that the AEMC has been monitoring. Indeed, we will use water collected from some of these sites as stressors in the lab to see if we can relate our prior results to this more "natural" perturbation. We are focusing efforts on distinguishing between competing theories of energy generation in this organism and on tracking electron flow in some of the inferred pathways. We will also extend our studies on the interaction of DvH with its syntrophic methanogens and derive a better view with how it operates within its community. Based on these studies will attempt to direct field studies to track the role of Dv-like species in affecting biogeochemical changes in the environment. We will also study further the evolution of regulation of key pathways in DvH by comparing and contrasting genomic data across the bacterial kingdom and functional genomic data from *Shewanella*, *Geobacter* and a number of other microbes. Ultimately, we will combine the discovered regulatory pathways and metabolic models of DvH into an integrated model of DvH physiology and begin to understand the evolutionary origins of this network by comparison to other environmental microbes.

61

Nitrate Stress Response in *Desulfovibrio vulgaris* Hildenborough: Whole-Genome Transcriptomics and Proteomics Analyses

Qiang He[1], Zhili He[1,2,6], Wenqiong Chen[3], Zamin Yang[1], Eric J. Alm[4,6], Katherine H. Huang[4,6], Huei-Che Yen[5,6], Dominique C. Joyner[4,6], Martin Keller[3,6], Jay Keasling[4,6], **Adam P. Arkin**[4,6], Terry C. Hazen[4,6], Judy D. Wall[5,6], and Jizhong Zhou[1,2,6]* (jzhou@ou.edu)

[1]Oak Ridge National Laboratory, Oak Ridge, TN; [2]University of Oklahoma, Norman, OK; [3]Diversa Corporation, San Diego, CA; [4]Lawrence Berkeley National Laboratory, Berkeley, CA; [5]University of Missouri, Columbia, MO; and [6]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

Sulfate reducing bacteria (SRB) are of interest for bioremediation with their ability to reduce and immobilize heavy metals. Nitrate, a common co-contaminant in DOE sites, is suggested to inhibit SRB via nitrite. Previous results indicate that nitrite is indeed inhibitory to the growth of *Desulfovibrio vulgaris*. However, growth inhibition by nitrate alone was also observed. In this study, growth and expression responses to various concentrations of nitrate were investigated using the Omnilog phenotype arrays and whole-genome DNA microarrays. Changes in the proteome were examined with 3D-LC followed by MS-MS analysis. Microarray analysis found 5, 50, 115, and 149 genes significantly up-regulated and 36, 113, 205, and 149 down-regulated at 30, 60, 120, and 240 min, respectively. Many of these genes (~50% at certain time points) were of unknown functions. By comparison to NaCl stress, transcriptional analysis identified changes specific to $NaNO_3$ stress. The hybrid cluster protein was among the highly up-regulated genes, suggesting its role in nitrate stress resistance with its proposed function in nitrogen metabolism. The up-regulation of phage shock protein genes (*pspA* and *pspC*) might indicate a reduced proton motive force and the repression of multiple ribosomal protein genes could further explain the growth cessation resulting from nitrate stress. A glycine/betaine transporter gene was also up-regulated, suggesting that $NaNO_3$ also constituted osmotic stress. Osmoprotectant accumulation as the major resistance mechanism was validated by the partial relief of growth inhibition by glycine betaine. Proteomics analyses further confirmed the altered expression of these genes, and in addition, detected increased levels of several enzymes (Sat, DvsB, and AprB) in the sulfate reduction pathway, indicative of the increased energy production during nitrate stress. In conclusion, excess $NaNO_3$ resulted in both osmotic stress and nitrate stress. *D. vulgaris* shifted nitrogen metabolism and energy production in response to nitrate stress. Resistance to osmotic stress was achieved primarily by the transport of osmoprotectant.

Many of the proteins that are candidates for bioenergetic pathways involved with sulfate respiration in *Desulfovibrio* spp. have been studied, but complete pathways and overall cell physiology remain to be resolved for many environmentally relevant conditions. In order to understand the metabolism of these microorganisms under adverse environmental conditions for improved bioremediation efforts, *Desulfovibrio vulgaris* Hildenborough was also used as a model organism to study stress response to nitrite, an important intermediate in the nitrogen cycle. Previous physiological studies demonstrated that growth was inhibited by nitrite and that nitrite reduction was observed to be the primary mechanism of detoxification. Global transcriptional profiling with whole-genome microarrays revealed a coordinated cascade of responses to nitrite in pathways of energy metabolism, nitrogen metabolism, oxidative stress

* Presenting author

response, and iron homeostasis. In agreement with previous observations, nitrite stressed cells showed a decrease in expression of genes encoding sulfate reduction functions in addition to respiratory oxidative phosphorylation and ATPase activity. Consequently, the stressed cells had decreased expression of ATP-dependent amino acid transporters and proteins involved in translation. Nitrite detoxification also appeared to shift the flow of reducing equivalents from oxidative phosphorylation to nitrite reduction. Increased demand for iron, resulting from these regulatory events and the chemical oxidation of available Fe^{2+}, likely contributed to iron depletion and the derepression of the Fur regulon.

62

Bacterial Nanowires: Novel Electron Transport Machines that Facilitate Extracellular Electron Transfer

Yuri A. Gorby[1]* (yuri.gorby@pnl.gov), Svetlana Yanina[1], Dianne Moyles[2], Matthew J. Marshall[1], Jeffrey S. McLean[1], Alice Dohnalkova[1], Kevin M. Rosso[1], Anton Korenevski[2], Terry J. Beveridge[2], Alex S. Beliaev[1], In Seop Chang[4], Byung Hong Kim[4], Kyung Shik Kim[4], David E. Culley[1], Samantha B. Reed[1], Margaret F. Romine[1], Daad A. Saffarini[3], Liang Shi[1], Dwayne A. Elias[1], David W. Kennedy[1], Grigoriy E. Pinchuk[1], Eric A. Hill[1], John M. Zachara[1], Kenneth H. Nealson[5], and Jim K. Fredrickson[1]

[1]Pacific Northwest National Laboratory, Richland, WA; [2]University of Guelph, Guelph, Ontario; [3]University of Wisconsin, Milwaukee, WI; [4]Korea Institute of Science and Technology, Seoul, Korea; and [5]University of Southern California, Los Angeles, CA

Coordinated transfer of electrons from electron donors to electron acceptors provides means for harvesting and transforming electrochemical potential energy into forms that sustain life. Electron transport components are truly 'global regulators' of cellular processes in the sense that energy flow influences all aspects of activity, response, and regulation and, as such, warrant scientific attention. Identifying electron transport components, characterizing the mechanisms of electron transduction, and relating electron transfer to bioenergetics are necessary to advance a systems level understanding of microorganisms. Dissimilatory metal reducing bacteria, such as *Shewanella oneidensis* strain MR-1 and *Geobacter sulfurreducens*, have enjoyed such scientific scrutiny since their discovery in the mid 1980's. Beyond obvious potential for these organisms to catalyze biogeochemical processes in natural environments, much of the research was driven toward developing a fundamental understanding of the processes that facilitate and limit electron transfer from bacteria to solid phases.

Although the controlling mechanisms of electron transfer remain poorly understood, electron acceptor availability under metal reducing conditions is typically a growth-limiting condition. Dissimilatory metal reducing bacteria produce electrically conductive appendages, which we call bacterial nanowires, in direct response to electron acceptor limitation. Nanowires produced by *S. oneidensis* strain MR-1, which served as our primary model organism, are functionalized by decaheme cytochromes MtrC and OmcA that are distributed along the length of the nanowires. Mutants deficient in MtrC and OmcA produce nanowires that were poorly conductive as determined by Scanning Tunneling Microscopy (STM). These mutants also differed from the wild type in their inability to reduce solid phase iron oxides, poor power production in a mediator-less microbial fuel cell, and failure to form complex biofilms at air-liquid interfaces.

Preliminary observations suggest that nanowires are also produced by other bacteria, including the oxygenic, phototrophic cyanobacterium *Synechocystis* PCC6803. This organism can produce highly conductive nanowires in response to excess light, which serves as the energy source to split water into

* Presenting author

protons and electrons, and limited CO_2, which otherwise serves as an electron sink during biomass production. Although additional work is needed to characterize the components of nanowires in other organisms, these results demonstrate that electrically conductive nanowires are not restricted to any single genus or even to particular metabolic guilds, such as dissimilatory metal reducing bacteria. Indeed, we hypothesize that nanowires are distributed throughout the bacterial world where they and serve as structures for efficient electron transfer and energy distribution. The electron carriers associated with nanowires in different metabolic guilds is likely to vary significantly due to differences in the redox couples utilized by the various organisms. High throughput methods such as those being proposed for new DOE biology user facilities would greatly facilitate the identification and analysis of these extracellular molecular machines. Further collaborative investigation into the complete composition of nanowires, mechanisms of electron flow through the wires, and interaction of nanowires between and among organisms in natural microbial communities is warranted in order to completely realize the implications of these structures in areas of alternative energy, carbon sequestration, bioremediation, and possibly pathogenicity and human health.

63

VIMSS Functional Genomics Core Research on Stress Response Pathways in Metal-Reducers

Aindrila Mukhopadhyay[1,2], Eric J. Alm[1,2], **Adam P. Arkin**[1,2,7], Edward E. Baidoo[1,2], Peter I. Benke[1,2], Sharon C. Borglin[1,2], Wenqiong Chen[1,3], Swapnil Chhabra[1,4], Matthew W. Fields[1,9], Sara P. Gaucher[1,4], Alex Gilman[1,2], Masood Hadi[1,4], Terry C. Hazen[1,2]* (tchazen@lbl.gov), Qiang He[1,5], Hoi-Ying Holman[1,2], Katherine Huang[1,2], Rick Huang[1,2], Zhili He[1,5], Dominique C. Joyner[1,2], Martin Keller[1,3], Keith Keller[1,2], Paul Oeller[1,3], Francesco Pingitore[1,2], Alyssa Redding[1,7], Anup Singh[1,4], David Stahl[1,8], Sergey Stolyar[1,8], Jun Sun[1,3], Zamin Yang[1,5], Judy Wall[1,6], Grant Zane[1,6], Jizhong Zhou[1,5], and Jay D. Keasling[1,2,7]* (keasling@Berkeley.edu)

[1]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov; [2]Lawrence Berkeley National Laboratory, Berkeley, CA; [3]Diversa, Inc., San Diego, CA; [4]Sandia National Laboratories, Livermore, CA; [5]Oak Ridge National Laboratory, Oak Ridge, TN; [6]University of Missouri, Columbia, MO; [7]University of California, Berkeley, CA; [8]University of Washington, Seattle, WA; and [9]Miami University, Oxford, OH

In collaboration with the Applied Environmental and Microbiology Core (AEMC), the Functional Genomics Core (FGC) has now fully optimized the generation of biomass for parallel experiments using the various genomics techniques. Outlined below are the most significant accomplishments from the different units of the FGC.

A. Microarray Analysis

Desulfovibrio vulgaris Hildenborough was used as a model organism to study a wide variety of stresses that are environmentally important. They include hyper-salinity, extreme pH conditions, Chromate, Nitrite, Nitrate and cold shock. Additionally, stationary phase physiology was also examined. Microarray analysis has now also been initiated for characterization of *D.vulgaris* mutants and the Δ*fur* mutant has been compared with wild type. Each transcriptome analysis has led to a vast repository of information regarding each stress. Several comparative stress response analyses are being conducted collaboratively between the FGC and the Computational Core (CC). Additionally, transcriptomics profiling has also been completed for several stresses in *S. oneidensis*. Initial studies

* Presenting author

have also been started for *G. metallireducens.* This sets the stage for important comparative studies that compare similar stress responses across different organisms.

B. Proteomics

a. Peptide tagging Quantitative proteomics using ICAT and ITRAQ strategies:

Methods were optimized for extensive separation of peptide pools so as to enable detection of a large number of peptides and corresponding proteins. Using a strong cation exchange coupled with reverse phase and tandem high resolution mass spectroscopy greater than 800 proteins were identified and their relative amounts quantified. The highly reproducible internal replicates allow stringent statistical analysis to be conducted on these data sets. This method had been applied to both oxygen stress and nitrate stress in *D.vulgaris.*

b. Comprehensive proteomics using 3D-LC/MS/MS

Using the powerful 3D separation technique and the high throughput mass spectroscopy at Diversa, Corp., a large number of proteins (40-60% of the proteome) were identified for several stress response comparisons. In this method spectral counting was used to estimate relative quantities of proteins. With proteomics data from several stress responses it now becomes possible to conduct a comparative proteomics analysis.

c. Study of Protein Complexes and Protein-Protein Interaction

Since proteins generally function within the cell through strong and weak interactions with other protein partners, it was considered necessary to characterize these complexes to increase our understanding of the functional proteomics picture. Two contrasting approaches for the isolation of protein complexes from *D. vulgaris* have been implemented. The endogenous approach involves *D.vulgaris* mutants containing an engineered tag that can be captured by affinity chromatography. Using lysates from these cells the tagged protein, in complex with its associated proteins, are captured and selectively eluted. In the exogenous approach, heterologously expressed His-tagged bait proteins from *D. vulgaris* are purified and coupled to affinity beads which are then incubated with *D. vulgaris* lysate to capture interacting proteins. Components of well characterized homologous protein complexes in *E. coli*, e.g. rpoB and rpoC have been used to validate methods. Several *D. vulgaris* proteins which have already been used in these strategies include Dnak, ClpX and CooX. Several ORFsgenes in Sulfate reducing bacteria are now being developed for similar studies.

C. Metabolomics

Fully developed methods have now been optimized by the FGC to study a vast majority of commonly encountered metabolites, which for methodological purposes are broadly broken down into molecules that are ideally resolved in positive ion mode or in negative ion mode. Given the scarcity of established methods in this area, our optimized strategies will be published as significant advancement in the field of metabolomics. A survey of metabolic extracts from *D. vulgaris* is being conducted as proof of concept. Noteworthy advance has also been made in the use of the Fourier Transform ion cyclotron resonance (FTICR) to detect and characterize metabolites based almost entirely on their exact mass. FTICR coupled Mass spectroscopy will enable the identification of hundreds of metabolites. In addition, fragmentation information can be used to characterize unknown peaks.

D. Integrating Genomics Data

Work from two separate stress response studies; namely Salt stress and Heat shock, were used in authoring integrated genomics manuscripts. The salt stress study was further complemented by phospholipid fatty acid analysis and several osmoprotection assays that led to a comprehensive model for salt stress in *D.vulgaris*. Following the model of rigorous checks for consistency and accuracy for all microarray data, an important collaborative project has been started between the FGC and CC to create a similar set of computational tools to analyze the Proteomics data. Data from the Oxygen stress and Nitrate stress are being used for this.

64

Comparative Analysis of Bacterial Gene Expression in Response to Environmental Stress

Eric J. Alm[1,2]* (ejalm@lbl.gov), Edward E. Baidoo[1,2], Peter I. Benke[1,2], Sharon C. Borglin[1,2], Wenqiong Chen[1,3], Swapnil Chhabra[1,4], Matthew W. Fields[1,9], Sara P. Gaucher[1,4], Alex Gilman[1,2], Masood Hadi[1,4], Terry C. Hazen[1,2], Qiang He[1,5], Hoi-Ying Holman[1,2], Katherine Huang[1,2], Rick Huang[1,2], Zhili He[1,5], Dominique C. Joyner[1,2], Jay D. Keasling[1,2,7], Martin Keller[1,3], Keith Keller[1,2], Aindrila Mukhopadhyay[1,2], Paul Oeller[1,3], Francesco Pingitore[1,2], Alyssa Redding[1,7], Anup Singh[1,4], David Stahl[1,8], Sergey Stolyar[1,8], Jun Sun[1,3], Zamin Yang[1,5], Judy Wall[1,6], Grant Zane[1,6], Jizhong Zhou[1,5], and **Adam P. Arkin**[1,2,7]

[1]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov; [2]Lawrence Berkeley National Laboratory, Berkeley, CA; [3]Diversa, Inc., San Diego, CA; [4]Sandia National Laboratories, Livermore, CA; [5]Oak Ridge National Laboratory, Oak Ridge, TN; [6]University of Missouri, Columbia, MO; [7]University of California, Berkeley, CA; [8]University of Washington, Seattle, WA; and [9]Miami University, Oxford, OH

The transcriptional response of bacterial species to environmental stress has been the subject of considerable research, fueled in part by the widespread availability of gene expression microarray technology. Previous studies have established the similarity of gene expression networks across a wide range of organisms, yet in these studies different experiments were performed on different species preventing a direct comparison. We have compiled a core set of 'standard' stressors including salt, pH, temperature, and oxygen and nitrite/nitrate levels and applied these stressors systematically to a phylogenetically diverse group of metal-reducing bacteria. We compare the expression patterns of orthologous genes and regulons in *Desulfovibrio vulgaris*, *Geobacter metallireducens*, and *Shewanella oneidensis* after exposure to these stressors. We observe that while the overall network may be conserved (genes in the same pathways have high correlations over all conditions), the response of the network to the same perturbations can be very different in different species (pathways may respond to the same stressor in different ways). Differences between species can arise from differential behavior of the same regulons and because 'orthologous' regulons may comprise different sets of (non-orthologous) genes, both of which may lead to insights in the ecological factors that shape gene expression.

* Presenting author

65

Evaluation of Stress Responses in Sulfate-Reducing Bacteria Through Genome Analysis: Identification of Universal Responses

J.D. Wall[1,7]* (wallj@missouri.edu), H.-C. Yen[1,7], E.C. Drury[1,7], A. Mukhopadhyay[2,7], S. Chhabra[3,7], Q. He[4], M.W. Fields[5,7], A. Singh[3], J. Zhou[4,7], T.C. Hazen[2,7], and **A.P. Arkin**[2,6,7]

[1]University of Missouri, Columbia, MO; [2]Lawrence Berkeley National Laboratory, Berkeley, CA; [3]Sandia National Laboratories, Livermore, CA; [4]University of Oklahoma, Norman, OK; [5]Miami University, Oxford, OH; [6]Howard Hughes Medical Institute, Berkeley, CA; and [7]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

The application of the toxic metal metabolism of the anaerobic sulfate-reducing bacteria to bioremediation of contaminated environments requires a broad understanding of the effects of environmental stresses on the organism. The model bacterium, *Desulfovibrio vulgaris* Hildenborough, for which the genome sequence has been fully determined, is being examined for its responses to a variety of stresses that may be expected to be encountered in natural/contaminated settings. We have examined the preliminary transcriptional data from ten treatments to learn whether there are general responses or common themes for responses to stresses by *D. vulgaris*. This anaerobe apparently does not have an ortholog encoding RpoS implicated in the universal stress response in γ-Proteobacteria. Interestingly genes predicted to be controlled by the global regulator Fur appear to be among the most frequently responsive in the genome. The transcriptional responses to increased concentrations of sodium and potassium overlapped strongly, as would be predicted. Curiously, it was not predicted that these salt responses would be shared by the response to reduced temperature. Also counter to our prediction, the response to nitrate was not a simple sum of the responses to sodium and nitrite. Further insights into general patterns of transcription during stresses will be discussed.

66

Cellular Responses to Changing Conditions in *Desulfovibrio vulgaris* and *Shewanella oneidensis*

M.E. Clark[1], Q. He[2], Z. He[2,7], K.H. Huang[3,7], E.J. Alm[3,7], X. Wan[1], T.C. Hazen[3,7], **A.P. Arkin**[3,4,5,7], J.D. Wall[6,7], J. Zhou[2,7], J. Kurowski[1], A. Sundararajan[1], A. Klonowska[1], D. Klingeman[2], T. Yan[2], M. Duley[1], and M. W. Fields[1,7]* (fieldsmw@muohio.edu)

[1]Miami University, Oxford, OH; [2]Oak Ridge National Laboratory, Oak Ridge, TN; [3]Lawrence Berkeley National Laboratory, Berkeley, CA; [4]University of California, Berkeley, CA; [5]Howard Hughes Medical Institute, Berkeley, CA; [6]University of Missouri, Columbia, MO; and [7]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

Temporal Transcriptomic Analysis of *Desulfovibrio vulgaris* Hildenborough Transition into Stationary-Phase Growth during Electron Donor Depletion

Desulfovibrio vulgaris was cultivated in a defined medium and biomass was sampled over time for approximately 70 h to characterize the shifts in gene expression as cells transitioned from exponential to stationary phase growth during electron donor depletion. In the context of *in situ* bioremediation, nutrient scarcity and/or depletion may be a common obstacle encountered by

* Presenting author

microorganisms due to the oligotrophic nature of most groundwater and sediment environments. In addition to temporal transcriptomics; protein, carbohydrate, lactate, acetate, and sulfate levels were measured. The microarray data was used for statistical expression analyses, hierarchical cluster analysis, and promoter element prediction. As the cells transitioned from exponential to stationary-phase growth a majority of the down-expressed genes were involved in translation and transcription, and this trend continued in the remaining time points. Intracellular trafficking and secretion, ion transport, and coenzyme metabolism showed more up-expression compared to down-expression as the cells entered stationary phase. Interestingly, most phage-related genes were up-expressed at the onset of stationary-phase. This result suggested that nutrient depletion may signal lysogenic phage to become lytic, and may impact community dynamics and DNA transfer mechanisms of sulfate-reducing bacteria. The putative feoAB system (in addition to other putative iron-related genes) was significantly up-expressed, and suggested the possible importance of Fe^{2+} acquisition under reducing growth conditions for sulfate-reducing bacteria. A large subset of carbohydrate-related genes had altered gene expression, and the total carbohydrate levels declined during the growth phase transition. Interestingly, the *D. vulgaris* genome does not contain a putative *rpo*S gene, a common attribute of the δ-*Proteobacteria* genomes sequenced to date, and other putative rpo factors did not have significantly altered expression profiles. The elucidation of growth-phase dependent gene expression is essential for a general understanding of growth physiology that is also crucial for data interpretation of stress-responsive genes. In addition, to effectively immobilize heavy metals and radionuclides via sulfate-reduction, it is important to understand the cellular responses to adverse factors observed at contaminated subsurface environments, such as the changing ratios of electron donors and acceptors. Our results indicated that genes related to phage, internal carbon flow, outer envelop, and iron homeostasis played important roles as the cells experienced electron donor depletion.

Deletion of a Multi-Domain PAS Protein Causes Pleiotropic Effects in *Shewanella oneidensis* MR-1

Shewanella oneidensis MR-1, a Gram-negative facultative anaerobe, can utilize a wide array of alternative electron acceptors during anaerobic respiration, and the ability to reduce soluble forms of heavy metals to insoluble forms makes it a potential candidate for bioremediation studies. Understanding the physiological responses of *S. oneidensis* to environmental stresses (e.g., nutrients, oxygen) is important for the assessment of potential impacts on metal-reducing activity. Here we describe the physiological role of a presumptive signal transduction protein in *Shewanella oneidensis* MR-1. The predicted ORF (SO3389) encoded a GGDEF, EAL, and two PAS domains. The deduced amino acid sequence was not closely related to previously described proteins, but presumptive proteins with similar domain architectures were observed in metabolically diverse microorganisms. An in-frame, deletion mutant was constructed (ΔSO3389), and the mutant displayed an extended lag period (30 h) when transferred from aerobic to anaerobic medium. The mutant was also defective in motility, cytochrome content, and was drastically defective in biofilm formation. These pleiotropic phenotypes were observed with multiple growth substrates. During the transition from aerobic to anaerobic conditions, the mutant was deficient in three c-type cytochromes (57, 33, and 20 kDa). In addition, mutant biofilms produced less carbohydrate compared to wild-type cells. Bacterial motility was affected only in aerobic conditions, and this result suggested that SO3389 was involved in O_2 responses and was important for anoxic and biofilm growth.

67

Probing Gene Expression in Single *Shewanella oneidensis* MR-1 Cells

X. Sunney Xie[1]* (xie@chemistry.harvard.edu), Paul Choi[1], Jie Xiao[1], Ji Yu[1], Long Cai[1], Nir Friedman[1], Jeremy Hearn[1], Kaiqin Lao[2], Luying Xun[3], Joe Zhou[4], Margaret Romine[5], and Jim Fredrickson[5]

[1]Harvard University, Cambridge, MA; [2]Applied BioSystems, Foster City, CA; [3]Washington State University, Pullman, WA; [4]Oak Ridge National Laboratory, Oak Ridge, TN; and [5]Pacific Northwest National Laboratory, Richland, WA

Our objective is to make real-time observations of gene expression in single live *Shewanella oneidensis* MR-1 cells with high sensitivity and high throughput. Available technology is only sufficient for the detection of gene expression at high expression levels. However, many important genes are expressed at low levels. New techniques are needed to probe gene expression that produces only a few protein molecules in a single cell, and to follow the expression in real time. Our efforts are summarized as follows:

Real time imaging of the production of single YFP molecules in a live cell

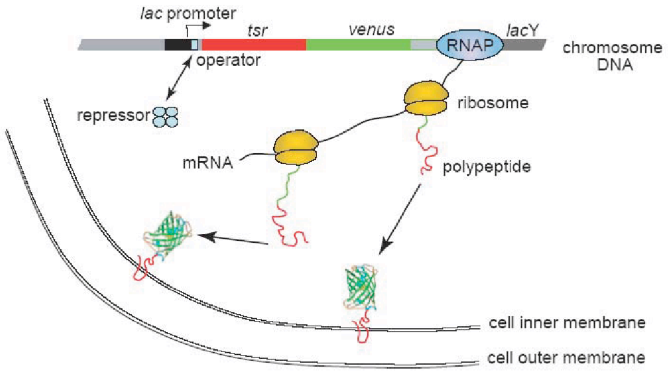
We have developed a reporter system for observing real-time production of single protein molecules in individual bacterial cells. A membrane-targeting sequence (tsr) was fused to the gene of fast maturing yellow fluorescent protein (YFP) under the control of a promoter on the chromosome DNA (Figure 1). Gene expression under a repressed condition generates membrane-localized YFP molecules that can be detected one at a time (Figure 2). We found that the protein molecules are produced in bursts and each burst originates from a stochastically transcribed single mRNA molecule. Protein copy numbers in the bursts follow an exponential distribution.

Real time monitoring of protein production in a live cell using β-galactosidase

We demonstrate another technique that allows measurements of low level protein expression in individual cells with single molecule sensitivity by taking advantage of the enzymatic properties of β-galactosidase (β-gal), the time-honored reporter for gene expression. β-gal can hydrolyze a wide range of synthetic substrates in addition to lactose, its native substrate. By hydrolyzing fluorogenic substrate a single enzyme molecule can produce a large number of fluorescent product molecules, as was first demonstrated by Rotman in 1961. However, live cell measurements have not been possible because efflux pumps on the cell membrane actively expel foreign organic molecules from the cytoplasm, inhibiting retention of fluorescent products in the cell.

To circumvent the efflux problem, we trap cells in closed microfluidic chambers, such that the fluorescent product expelled from a single cell can accumulate in the small volume of a chamber, recovering the fluorescence signal due to enzymatic amplification.

Figure 1. Detection of single membrane-immobilized YFP molecules that are generated under a repressed condition.



* Presenting author

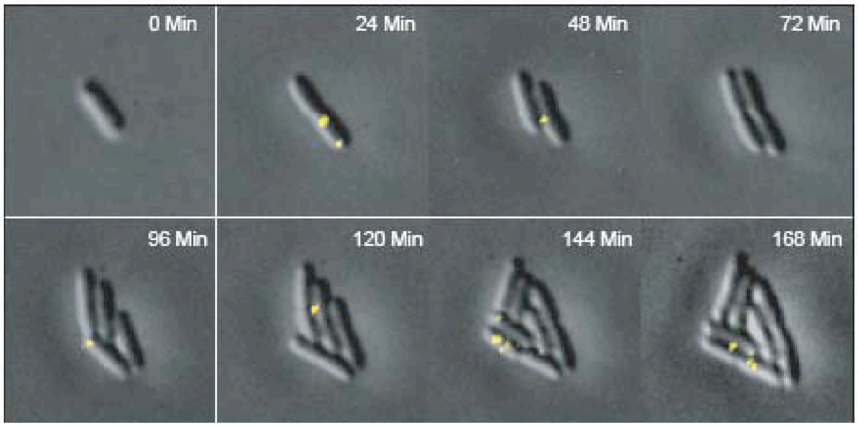


Figure 2. Sequence of fluorescent images of single molecules of fast maturing and membrane-immobilized YFP (yellow) overlaid with bacterial images.

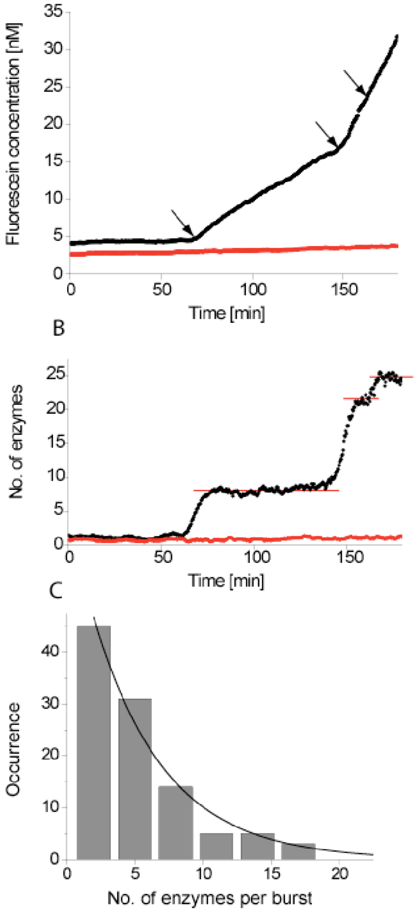


Figure 3. (A) Abrupt changes in fluorescence signal of a chamber with and with out dividing cells (flat curve). (B) The time derivative of the trace in A. (C) Exponential distribution of copy number of β-gal per expression burst.

We observed discrete changes of the slopes of fluorescence signal (Figure 3A) indicating that protein production occurs in bursts (Figure 3B), again with the number of molecules per m-RNA following an exponential distribution (Figure 3C). We show that the two key parameters of protein expression, the burst size and frequency, can be either determined directly from real time monitoring of protein production or extracted from a measurement of the steady-state copy number distribution in a population of cells with a simple theoretical model.

These studies not only provide new methods for quantification of low-level gene expression, but also yields quantitative understanding of the working of transcription and translation in live cells.

Real time RT-PCR for sensitive mRNA quantification

In collaboration with ABI, we use real time RT-PCR to quantitatively measure low mRNA copy numbers inside single bacterial cells. The expression of the reporter gene at the mRNA level can be correlated with that at the protein level.

Compared to the existing methods for characterization of gene expression, such as DNA microarrays and mass spectrometry, the above three methods allow high sensitivity and measurements of gene expression profiling in single live cells. It enables studies of gene expression at the uncharted low expression levels, providing a complete picture of global gene expression.

Library construction of Shewanella oneidensis MR-1

We have developed an efficient system for targeted insertions of YFP-gene translational fusions into the *Shewanella oneidensis* MR-1 chromosome using the Gateway technology. Our initial library of over a hundred sequence-confirmed strains includes cytochrome and biofilm-related genes. We are applying the above new techniques to study gene expression of *Shewanella oneidensis.*

* Presenting author

68

Phosphoproteome of *Shewanella oneidensis* MR1

Gyorgy Babnigg[1]* (gbabnigg@anl.gov), Cynthia Sanville-Millard[1], Carl Lindberg[1], Sandra L. Tollaksen[1], Wenhong Zhu[2], John R. Yates III[2], Jim K. Fredrickson[3], and **Carol S. Giometti**[1]

[1]Argonne National Laboratory, Argonne, IL; [2]Scripps Research Institute, La Jolla, CA; and [3]Pacific Northwest National Laboratory, Richland, WA

Protein phosphorylation plays an important role in the regulation of cell physiology in both prokaryotes and eukaryotes. Although bacterial protein phosphorylation on histidine residues is well known to be involved in signal transduction, phosphorylation of serine, threonine or tyrosine residues has only recently been described. We are using a suite of proteomics tools, including affinity chromatography, two-dimensional gel electrophoresis (2DE), Western blotting, tryptic peptide mass analysis, and phosphopeptide characterization to identify phosphoproteins expressed by *Shewanella oneidensis* MR-1 cells grown under different conditions and to determine whether or not serine, threonine, or tyrosine phosphorylation events are involved in the regulation of *S. oneidensis* MR-1 metabolism. By probing 2DE patterns of whole cell lysate patterns with antibodies directed against specific phosphorylated-amino acids, proteins containing phosphoserine, phosphotyrosine, and phosphothreonine have been detected. In a comparison of cells grown aerobically and with oxygen limitation (suboxic), differential expression of phosphoproteins has been observed (Figure. 1).

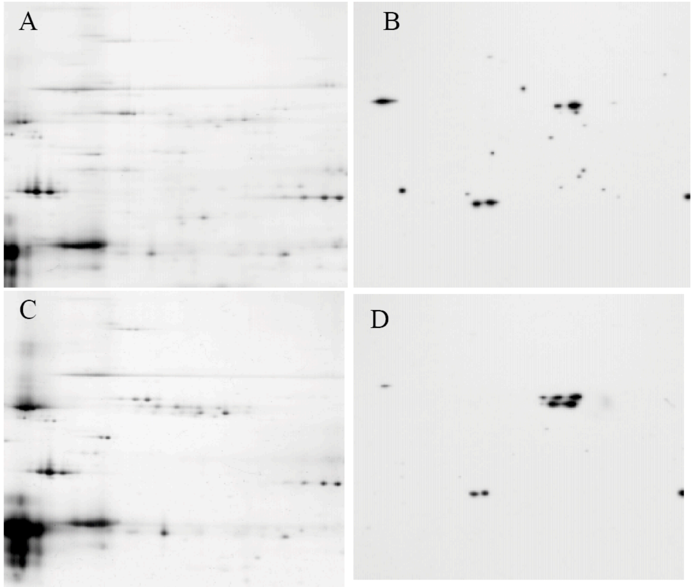


Figure 1. 2DE patterns of MR-1 phosphoproteins isolated by affinity chromatography and detected by silver stain (A,C) or by reaction with anti-phosphoserine antibody (B,D). Cells were grown aerobically (A,B) or with limited oxygen (C,D).

Sixteen phosphoproteins have been identified by tryptic peptide mass analysis using LC-MS/MS; identified proteins include GGDEF domain protein, translation elongation factor Tu, and formate acetyltransferase. Phosphorylated proteins have also been enriched from cell lysates using immobilized metal affinity chromatography prior to analysis by 2DE or LC-MS/MS. The isolation of phosphoproteins by affinity chromatrography is also enabling characterization by Fourier transform MS/MS to identify the actual sites of phosphorylation and to determine whether or not there is site-specific phosphorylation in response to different growth conditions. Such characterization is of particular interest in the case of formate acetyltransferase (Figure 2), since this protein has more phosphorylated forms under suboxic than under aerobic growth

* Presenting author

conditions, suggesting a regulatory function (Figure 2). This work will be extended to the study of cells harvested from the *Shewanella* Federation chemostat experiments in order to follow the phosphorylation events through transition between different growth conditions.

This work is supported by the U.S. Department of Energy, Office of Biological and Environmental Research, Genomics:GTL Program, under contract number W-31-109-ENG-38.

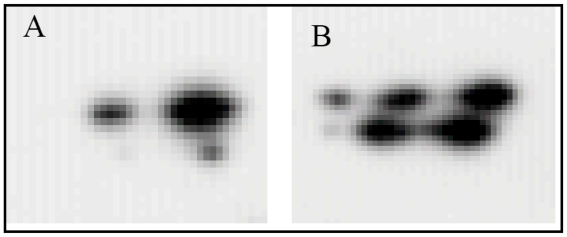


Figure 2. 2DE pattern of formate acetyltransferase from MR-1 cells grown aerobically (A) or with limited oxygen B); detected using anti-phosphoserine.

69

The *Shewanella* Federation: Functional Genomic Investigations of Dissimilatory Metal-Reducing *Shewanella*

James K. Fredrickson[1]* (Jim.Fredrickson@pnl.gov), Margaret F. Romine[1], Carol S. Giometti[2], Eugene Kolker[3], Kenneth H. Nealson[4], James M. Tiedje[5], and Jizhong Zhou[6,11], Monica Riley[7], Shimon Weiss[8], Christophe Schilling[9], and Timothy S. Gardner[10]

[1]Pacific Northwest National Laboratory, Richland, WA; [2]Argonne National Laboratory, Argonne, IL; [3]BIATECH, Bothell, WA; [4]University of Southern California, Los Angeles, CA; [5]Michigan State University, East Lansing, MI; [6]Oak Ridge National Laboratory, Oak Ridge, TN; [7]Marine Biological Laboratory, Woods Hole, MA; [8]University of California, Los Angeles, CA; [9]Genomatica, Inc., San Diego, CA; [10]Boston University, Boston, MA; and [11]University of Oklahoma, Norman, OK

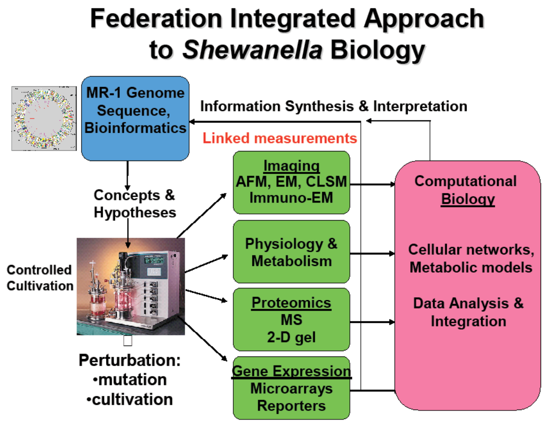
Shewanella oneidensis MR-1 is a motile, facultative γ-*Proteobacterium* with extensive metabolic versatility with regards to electron acceptor utilization; it can utilize O_2, nitrate, fumarate, TMAO, DMSO, Mn, Fe, and S^0 as terminal electron acceptors during anaerobic respiration. The ability to effectively reduce polyvalent metals including solid phase Fe and Mn oxides and radionuclides such as uranium and technetium as well as to function as a catalyst in mediator-less microbial fuel cells has made *Shewanella* an excellent model organism for understanding biogeochemical cycling of metals and anaerobic electron transport pathways. Complete sequencing of the MR-1 genome has enabled the application of high throughput functional genomics methods for measuring gene and protein expression. The *Shewanella* Federation (SF), a collaborative scientific team, is applying these approaches to achieve a system-level understanding of how MR-1 regulates energy and material flow and to utilize its versatile electron transport system to reduce metals and transfer electrons to electrode surfaces. The SF has developed an integrated approach to *Shewanella* functional genomics that capitalizes on the relative strengths, capabilities, and expertise of the various members. A major emphasis has been placed on integrating prediction and experiment in an iterative fashion to unravel the network structure that controls the flow of energy and materials through cells. SF members share information, resources and collaborate on projects that range from a few investigators focused on a defined topic to more complex "Federation-level" experiments that utilize combined SF capabilities to address more global scientific questions. The SF is organized into integrated working groups focused on: (1) the functional genomics of energy flow and electron transport regulation; (2) the characterization and modeling of metabolic and regulatory networks; (3) the comparative genomic and physiologic analyses of multiple *Shewanella* species to develop an evolutionary model for

* Presenting author

Shewanella; and (4) the genetic and functional characterization of *c*-type cytochromes to determine their relative roles in respiratory metabolism. In support of SF science objectives, efforts were recently initiated to develop an integrated knowledge and data sharing resource. The SF also has informal collaborations with a number of independent Genomics:GTL projects, many of which are focused on developing new experimental and computational capabilities.

The respiratory versatility of *Shewanella* is believed to be benefited by the remarkably diverse and complex electron transport system and a relatively large number of *c*-type cytochromes. In spite of substantial effort, however, the details of MR-1's electron transport system and the mechanisms by which it transfers electrons to metals and electrode surfaces remain unknown. Recent results have confirmed the role of extracellular multi-heme *c*-type cytochromes and have localized these cytochromes to specific structures termed nanowires. Additional work is needed to understand the composition and function of these extracellular molecular machines that appear to be responsible for electron transfer to metals as well as electrode surfaces. Even less is known regarding the global networks in this organism that allow it to respond to changing environmental conditions and regulate carbon and energy flow. Applying DNA microarray and proteome technologies, coupled with controlled cultivation and detailed analyses of physiology and cell composition and modeling has great potential to achieve a system-level understanding of how MR-1 regulates energy and material flow and to utilize its remarkably versatile electron transport system to reduce metals. Integrated SF experiments typically take the generic form as illustrated in Figure 1.

Figure 1. *Shewanella* Federation model for integrated experimentation



In this model, the MR-1 genome annotation is continuously updated using a combination of bioinformatics tools, experimental results and manual inspection. Controlled cultivation and mutagenesis are used to perturb the biology in a very controlled manner, while parallel analyses of gene and protein expression patterns as well as measurements of metabolites and cell physiology are used to assess how MR-1 responds to controlled perturbations. This information is, in turn, used to identify regulatory networks and to test hypotheses regarding metabolism and to revise the metabolic model of MR-1. Various imaging technologies are used to assess the physical state of the cell and to obtain insight into composition. In addition to providing specific system-level insights into the biology of *Shewanella*, the results from collaborative experiments also serve as a general resource for addressing broader genomics questions. For example, the collective microarray and proteome data from various experiments has been used to probe the expression of hypothetical genes and small proteins in the MR-1 genome and to provide experimental evidence for the general metabolic pathways predicted by the genome sequence. While the SF has emphasized large integrated experiments, there have also been numerous subprojects involving multiple SF participants that address a variety of subtopics ranging from understanding the effects of temperature and ionizing radiation on cell physiology and gene expression to the role of specific global regulators in controlling various subnetworks involved in aerobic and anaerobic metabolism. As a critical step for studying protein-ligand interactions using phage-display, yeast two-hybrid systems, the majority of *Shewanella* genes were cloned and verified. The clone set is a valuable resource for further investigating gene functions, regulatory networks and molecular machinery in *S. oneidensis* MR-1.

* Presenting author

The SF holds biannual meetings that are rotated among the various partners. These meetings include PIs, collaborators, staff and students and are used to review and plan collaborative research and to discuss a variety of technical issues ranging from data sharing mechanisms to growth and mutagenesis protocols. These meetings are open and often include non-DOE funded researchers who have an interest in the biology of *Shewanella*. SF members share experimental data, protocols and materials (mutants, reporter gene constructs, clones etc.) and maintain a collective annotation of the MR-1 genome that is based on input and data from multiple sources. Future directions of the SF include investigations into the population and community biology of *Shewanella* to begin to link genomics to populations, communities and evolution.

70

An Integrated Knowledge Resource for the *Shewanella* Federation

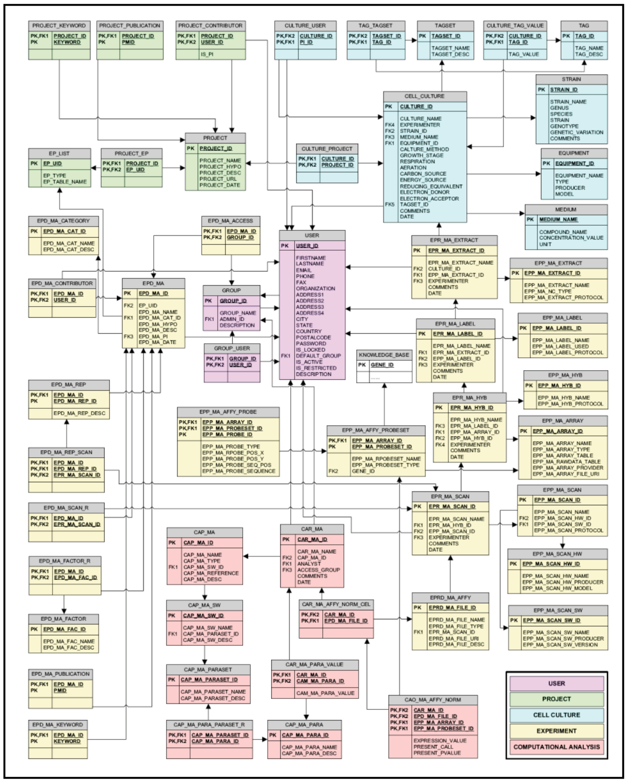
Nagiza F. Samatova[1,*] (samatovan@ornl.gov), Nicole Baldwin[1], Bing Zhang[1], Buyng-Hoon Park[1], Denise Schmoyer[1], Yuri Gorby[2], Grigoriy E. Pinchuk[2], Timothy S. Gardner[3], Mary S. Lipton[2], Samuel Purvine[2], Angela D. Norbeck[2], Margaret F. Romine[2], Gyorgy Babnigg[4], Carol S. Giometti[4], James K. Fredrickson[2], and **Ed Uberbacher**[1]

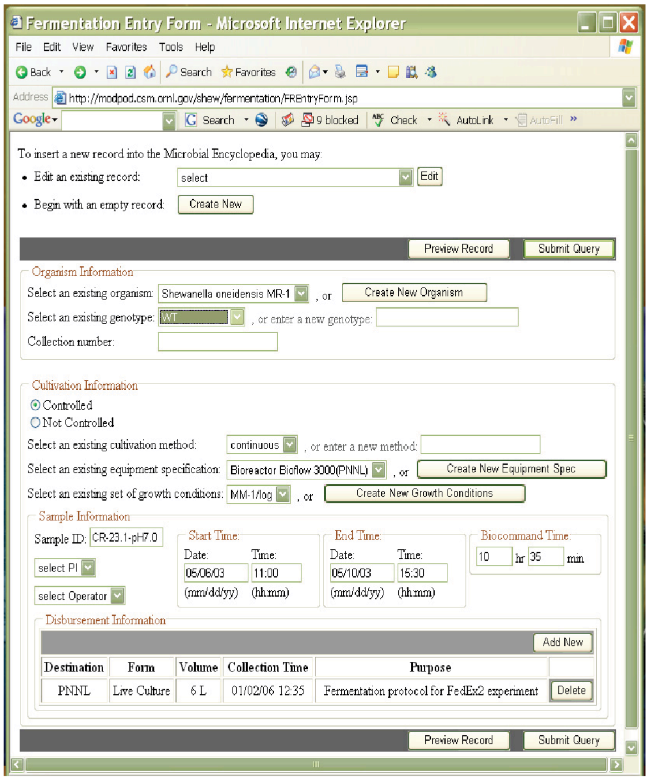
[1]Oak Ridge National Laboratory, Oak Ridge, TN; [2]Pacific Northwest National Laboratory, Richland, WA; [3]Boston University, Boston, MA; and [4]Argonne National Laboratory, Argonne, IL

The *Shewanella* Federation (SF) represents a distributed group of investigators formed to understand the metabolic potential of *Shewanella* species, particularly with relevance to metal reduction and bioremediation applications. The Federation is generating large volumes of data of many different types, however, current data management strategies primarily rely on localized solutions and *ad hoc* data exchange procedures. The lack of integrative bioinformatics solutions is an impediment to bringing this major DOE investment to its full potential. Based on highest priority needs presented by many individual SF researchers, this project aims to construct a data and knowledge integration environment that will allow investigators to query across the individual research domains, link to analysis applications, visualize data in a cell systems context, and produce new knowledge, while minimizing the effort, time and complexity to participating laboratories. Specifically, are major goals are: (1) to develop strategies for capturing and integrating diverse data types into common data models that support systems biology investigation, (2) to develop tools and processes to catalog and retrieve high-throughput data from warehoused and non-local data storage, (3) to construct a data and knowledge base that integrates gene, protein, expression and pathway-level knowledge, and (4) to incorporate interfaces for navigation and visualization of the multi-dimensional data produced.

During the initial three months of the project the following progress has been made. We have configured and built a server hardware and software infrastructure to provide efficient and reliable (24/7; 100% reliability; zero data loss) future access to the data generated by SF. We collected sample data files and associated meta-data including those for experimental protocols, raw experimental data, pre-processed and computationally analyzed data from the majority of SF sites to help with system design. Based in part on this input, we designed a comprehensive relational schema, compliant with community accepted data standards, that currently is capable of capturing fermentation, microarray, proteome, and interactome data. This high-level schema design captures information about Projects, Cell Culture, Experiments (fermentation, microarray, proteomics, MS pull-downs, etc.), Computational Analyses, and User(s) (figure, next page). For some sites, where local relational schemas are available, the proper schema mapping strategies have been developed. In

[*] Presenting author




Fermentation Entry Form - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address
Google
Search 9 blocked Check AutoLink AutoFill

Address: http://modpod.csm.ornl.gov/shew/fermentation/FREntryForm.jsp

To insert a new record into the Microbial Encyclopedia, you may:

- Edit an existing record: select Edit
- Begin with an empty record: Create New

Preview Record Submit Query

Organism Information

Select an existing organism: Shewanella oneidensis MR-1, or Create New Organism

Select an existing genotype: WT, or enter a new genotype:

Collection number:

Cultivation Information

● Controlled

○ Not Controlled

Select an existing cultivation method: continuous, or enter a new method:

Select an existing equipment specification: Bioreactor Bioflow 3000(PNNL), or Create New Equipment Spec

Select an existing set of growth conditions: MM-1/log, or Create New Growth Conditions

Sample Information

Sample ID: CR-23.1-pH7.0

select PI

select Operator

Start Time: Date: 05/06/03 Time: 11:00 (mm/dd/yy) (hh:mm)

End Time: Date: 05/10/03 Time: 15:30 (mm/dd/yy) (hh:mm)

Biocommand Time: 10 hr 35 min

Disbursement Information

Add New

Destination	Form	Volume	Collection Time	Purpose	
PNNL	Live Culture	6 L	01/02/06 12:35	Fermentation protocol for FedEx2 experiment	Delete

Preview Record Submit Query

addition, we designed a schema for publicly available data sources and integrated this schema with the one for SF experimental data and meta-data. We wrote various tools for data ingest from these public databases, data parsing, and upload to the *Shewanella* knowledgebase including sequence databases (GenBank, TIGR, RefSeq, UniProt, InterPro, Pfam, ProDom, SMART), structure databases (PDB, COILS, SOSUI, PROSPECT), pathway databases (KEGG), and protein interaction databases (STRING, 3DID, DIP). Each of these data sources can be queried through a common interface that supports both simple and advanced search capabilities suitable for both browsing and for complex queries. To accept data and meta-data from various researchers in SF, we developed a web-based system for users to upload the information relevant to their experiments, and automatically update the database, thus making this information searchable across all the previously entered information. The system currently supports fermentation experiments and is being extended to other protocols (see figure, above). The system provides the capability for SF sites to edit and annotate the information, to save experimental protocol description in Excel format, to restrict the choices using controlled vocabularies, and to validate entered fields. The system is based on XML, XSD, and JSP/Bean technologies. On the analysis side, we developed tools for quantification of data from either stable isotope labeling or label-free LC-MS/MS shotgun proteomics. The tools will be freely available from http://MSProRata.org. In addition, we applied our protein-protein interaction prediction tools to *Shewanella* and are making the results of these predictions available through the *Shewanella* knowledgebase http://modpod.csm.ornl.gov/shew.

Research sponsored by the DOE Genomics:GTL Program.

* Presenting author

71

Physiological Characterization of Genome-Sequenced *Shewanella* Species

Radu Popa, Rodica Popa, Anna Obraztsova, A. Hsieh, and **Kenneth H. Nealson*** (knealson@usc.edu)

University of Southern California, Los Angeles, CA

The genus *Shewanella* consists of a widely distributed group of facultatively anaerobic bacteria that are renowned for their ability to reduce many different electron acceptors and in particular for their ability to reduce solid phase iron and manganese oxides. *S. oneidensis* MR-1 was the first of this genus to be sequenced and annotated, followed by a number of other strains through DOE's Joint Genome Institute in cooperation with the *Shewanella* Federation (SF), providing a great opportunity for combined comparative physiology and genomic analyses. That is, to ask, among other questions, the extent to which the physiology of a given bacterium can be predicted by analysis of its genome, and if not, what must we learn to be able to accurately do so. A necessary step in this co-analysis is the physiological analysis of the sequenced strains. This report includes a summary of physiological characterization of several sequenced species and strains of *Shewanella*. A companion abstract (J. Tiedje et al.) describes the initial results of genomic and proteomic analyses of a subset of the sequenced strains.

The ability to respire different electron donors was tested via the Biolog™ phenotype array system. This system utilizes 96-well plates, in which each well contains a different energy source. If a substrate is respired, a dye (formazan) is reduced, producing a blue color that is scored using a plate reader. Thus, the system scores respiratory ability, but not necessarily growth. Two different Biolog™ plates were utilized, providing a total of approximately 190 different substrates. Of these, approximately 10 were respired by all 7 strains, and another 15 were respired by 6 of the eight strains tested. Some strains, such as MR-1 utilized only 22 different substrates, while others had a much wider substrate range, up to 41 different compounds, including many hexoses and complex molecules.

The ability to grow on substrates that can be respired was tested on all substrates that were positive for respiration. This work is in progress, but relates to the issue of whether a given substrate might be susceptible to oxidation, yet not utilized as a substrate for growth. For strain MR-1, a surprisingly large number of substrates that were respired were not capable of supporting growth in the standard minimal medium.

The ability to grow anaerobically utilizing different electron acceptors was tested on a variety of different electron acceptors. This work is also in progress, and involves the screening of each strain for the ability of several substrates that were known to support growth aerobically to grow anaerobically on various electron acceptors. While the physiological mechanisms are not yet elucidated, it is clear that many strains utilize some electron donors quite well for some electron acceptors, but not for others. These results suggest that the internal electron transport webs of MR-1, and other shewanellae, are more complex than has been appreciated. A current objective of the *Shewanella* Federation is to elucidate the role of various *c*-type cytochromes in anaerobic respiration, in *S. oneidensis* MR-1, via a combination of bioinformatics and experimental analyses. This will greatly facilitate cross-genome comparisons and testing of hypotheses regarding the role to specific cytochrome homologs across the different strains.

All in all, the strains so far analyzed exhibit a wide range of physiological diversity, consistent with observations of genetic diversity at the genome scale, thus providing ample fodder for additional detailed genomic/physiology analyses.

* Presenting author

72

The Use of Microbial Fuel Cells (MFCs) for the Study of Electron Flow to Solid Surfaces: Characterization of Current Producing Abilities of Mutants of *Shewanella oneidensis* MR-1, of Other Strains and Species of *Shewanella*

Orianna Bretschger[1]*, Byung Hong Kim[2], In Seop Chang[2], Margaret F. Romine[3], Jim K. Fredrickson[3], Yuri A. Gorby[3], Samantha B. Reed[3], David E. Culley[3], Soumitra Barua[4], and **Kenneth H. Nealson**[1]

[1]University of Southern California, Los Angeles, CA; [2]Korea Institute of Science and Technology, Seoul, Korea; [3]Pacific Northwest National Laboratory, Richland, WA; and [4]Oak Ridge National Laboratory, Oak Ridge, TN

Shewanella oneidensis MR-1 is a Gram negative facultative anaerobe capable of utilizing a broad range of electron acceptors, including several solid substrates. In addition, it has been known for many years that MR-1 can catalyze current production in microbial fuel cells (MFCs) without the addition of electron shuttles. We have examined a number of mutants of strain MR-1 for their ability to produce current, in an effort to determine whether the pathway of electron flow to solid metal oxides might be the same, or similar to, that used for current production. If so, it may be possible to use the MFC as a quantitative measure of electron flux: essentially as a proxy for the ability to reduce solid substrates. We summarize herein the results of MFC screening of mutants altered in both structural and regulatory genes.

Our microbial fuel cell is a two chambered cell separated by a proton-permeable membrane. On the anode side, *S. oneidensis* MR-1 is contained under anaerobic conditions, where it catalyzes the oxidation of substrate, producing electrons that flow through the anode electrode and across an external circuit. The protons produced by MR-1 diffuse through the membrane to the cathode side. A platinum catalyst is used at the cathode to convert the electrons, protons and molecular oxygen to water.

Mutants defective in various cytochromes, as well as various regulatory elements has been made by members of the *Shewanella* Federation (SF). These mutants are targeted deletions constructed by either homologous cross-over using host-encoded recombinases (PNNL group), or by introduced phage cre-lox recombinases (ORNL group). These mutants display a wide range of properties, but the over-riding feature seen is that mutants that inhibit the ability of MR-1 to reduce solid phase iron also inhibit the ability to produce current in the MFC, and to a similar degree, suggesting that the mechanisms involved in both processes require similar regulatory and structural components. In particular, we note that mutants in the *mtrA, mtrB, mtrC* or *omcA* structural genes are always deficient in both iron reduction and current production.

In addition, a number of *Shewanella* species and strains were screened for their ability to produce current. These included, in addition to *S. oneidensis* MR-1, 7 strains that were recently sequenced by the JGI. Of interest in this study was the strain 0S217, *S. denitrificans*, which is lacking the cassette of genes (*mtrA, B, and C, and omcA*) that are believed responsible for current production and solid iron reduction in the other *Shewanella* strains. All of the strains, including OS217 produced comparable current, suggesting that a different mechanism must be utilized for both metal reduction and current production by this organism.

* Presenting author

In addition to serving as a platform for electron flux studies, the MFC was also used for the analysis of nutrients during current production. Monitoring of reactants and products during current production allowed real time assessment of metabolic flow as related to current production.

73

Characterization of *c*-Type Cytochromes in *Shewanella oneidensis* MR-1

Margaret F. Romine[1]* (margie.romine@pnl.gov), Samantha B. Reed[1], Soumitra Barua[2,5], Samuel O. Purvine[1], Alex S. Beliaev[1], Haichun Gao[2,5], Zamin Yang[2,5], Yunfeng Yang[2,5], David E. Culley[1], David W. Kennedy[1], Yuri Londer[3], Tripti Khare[3], Sandra Tollaksen[3], Claribel Cruz-Garza[4], Jun Li[6], Mandy Ward[6], Jizhong Zhou[2,5], Frank Collart[3], Mary S. Lipton[1], James M. Tiedje[4], Carol S. Giometti[3], and Jim K. Fredrickson[1]

[1]Pacific Northwest National Laboratory, Richland, WA; [2]Oak Ridge National Laboratory, Oak Ridge, TN; [3]Argonne National Laboratory, Argonne, IL; [4]Michigan State University, East Lansing, MI; [5]University of Oklahoma, Norman, OK; and [6]Johns Hopkins University, Baltimore, MD

Shewanella oneidensis MR-1 is a facultative aerobe that is capable of using fumarate, nitrate, nitrite, dimethylsulfoxide (DMSO), thiosulfate, trimethylamine oxide (TMAO), S°, and a variety of solid-phase and complexed metals including Fe(III) oxides, Mn(IV) oxides, Tc(VII), U(VI), and V(V) to drive respiration. Such a unique respiratory versatility of dissimilatory metal reducing bacteria like MR-1 is largely due to the abundance of *c*-type cytochromes that are revealed upon sequencing of their genomes. The predicted localization of a subset of these cytochromes to the outer membrane is thought to enable these organisms to utilize insoluble electron acceptors such as metal oxides and S°. Thus far four *c*-type cytochromes, OmcA, MtrC, MtrA, and CctA have been linked to respiration of Fe(III) and Mn(IV) by *Shewanella sp.* and an additional 10 are predicted to be necessary for respiration of oxygen, fumarate, nitrate, nitrate, DMSO, TMAO, or S° by MR-1. An additional cytochrome *c*, CymA, is required for respiration of Fe, fumarate, DMSO, nitrite, and nitrate, but not TMAO. The roles of the remaining predicted cytochromes (nearly 75% of the total number) are have not yet been elucidated. This observation confirms that the mechanisms of electron transport in *Shewanella* remain poorly understood. Consequently, the *Shewanella* federation has developed a strategy to explore this metabolism further using an integrated approach that involves targeted gene knock-outs as well as comparative genomics, physiology, and gene/protein expression patterns. Summarized herein is our progress in determining the roles of *S. oneidensis* MR-1 *c*-type cytochromes in cellular respiration.

Prediction of Genes Encoding *c*-Type Cytochromes. Computational analysis of the MR-1 genome sequence revealed that the deduced amino acid sequences from 71 genes contain the signature CXXCH heme *c* binding motif. Functions of orthologs in other bacteria, the occurrence of conserved domains, and the presence of an expected sec leader peptide were used to narrow the number of predicted cytochromes to 44. Comparison to proteins deduced from genome sequences of 10 other *Shewanella* sp. enabled us to determine that SO3141 is degenerate, requiring 4 frameshifts to reconstruct the proper reading frames to produce the expected intact outer membrane decaheme cytochrome *c*. Furthermore, it was revealed that the 3' and 5' end of genes SO4570 and SO4569, encoding a cytochrome and NfrC-like FeS protein, respectively, are replaced by 6 repeats of CAAGTGGTA. A third gene, SO3623 encodes a split tetraheme flavocytochrome *c* and is intact, but the upstream gene encoding the flavin subunit is interrupted by an IS element. Consequently, it is unlikely that these proteins participate in respiratory processes in MR-1 without genomic rearrangement. In summary we have identified 41 genes that are predicted to encode *c*-type cytochromes.

* Presenting author

Validation of *c*-Type Cytochrome Predictions. The occurrence of a CXXCH motif and a signal peptide is not sufficient to confidently determine that a protein is a *c*-type cytochrome since a large number of functionally unrelated proteins also possess a CXXCH motif. The covalently attached heme characteristic of *c*-type cytochromes can easily be detected by staining proteins separated by SDS-PAGE with a chemiluminescent ELISA substrate and therefore used as a method to validate our predictions and to map their mobility in acrylamide gels to facilitate their identification in ongoing 2D PAGE-based differential protein expression studies. We have successfully overexpressed, in *Escherichia coli,* 15 of the proteins predicted to bind 4 or fewer hemes and demonstrated that they bind heme in 1D acrylamide gels. An additional 3 proteins were expressed but not soluble. In addition to location mapping in 2D gel separation systems, these protein preparations will also be characterized by AMT mass tag technology to identify signature peptides that uniquely identify these proteins in cellular protein extracts.

Conditions that Promote Expression of Predicted *c*-Type Cytochromes. The identification of conditions that uniquely enable expression of RNA or protein from these genes provide useful clues of their function and determine the conditions necessary for detecting aberrant phenotypes in mutants. Microarray analyses of MR-1 cells grown with fumarate were evaluated for changes in gene expression after a shift to 10 alternate electron acceptors (Beliaev, et al. 2005). Results revealed a surprisingly widespread induction of cytochrome *c* genes with thiosulfate and conversely the paucity of genes induced by DMSO. Many of the genes induced include those which are functionally uncharacterized suggesting that growth or reduction of thiosulfate and related sulfur containing compounds should be further investigated.

AMT tag proteome analysis of MR-1 extracts collected from aerobic, suboxic, and anaerobic cultures using fumarate as the electron acceptor have confirmed expression of 22 of these genes, including several encoded by genes whose expression did not change significantly in microarray analyses. The combined results from these expression experiments allow us to distinguish the cytochromes that are required for multiple respiratory metabolisms from those that are unique to one or a few types.

Targeted Deletion of *c*-Type Cytochromes. Targeted deletion of all but 5 of the predicted intact cytochrome *c* encoding genes have been successfully constructed by either homologous cross-over with host-encoded recombinases (PNNL) or with introduced phage *cre-loxP* recombinases (ORNL). Each mutant was tagged with a unique bar code to facilitate tracking individual strains in planned competitive growth studies. Preliminary analyses on the ability of these mutants to grow in microtiter plates with different electron acceptors revealed 5 mutants with growth defects with nitrate. Surprisingly the Δ*napB* mutant grew better in LB/nitrate medium than wild type cells, while a Δ*napA* mutant could not grow at all. Nitrite, which is toxic, accumulated after 12 hours of growth by the WT strain only. We hypothesize that in the absence of *napB*, an alternative cytochrome c supplies electrons to *napA* for subsequent reduction of nitrate. Removal of nitrite by this alternative pathway is more efficient thereby enabling cells to attain a higher biomass yield. Mutants in the high-affinity cbb_3 cytochrome oxidase components exhibit a defect in both O_2 and TMAO electron acceptors suggesting a role for this complex in both suboxic and anaerobic respiratory processes. Defects in the reduction of Mn(IV) relative to WT MR-1 was evident in 10 different mutants suggesting a complex network of electron transfer reactions. Mutants were also evaluated anaerobically for energy taxis to Fe(III)-citrate, nitrate, nitrite, TMAO, DMSO, and fumarate using swarm plate assays. Initial test results suggest that 2 mutants were defective in all 6 assays, while others showed defects for only selected substrates providing new clues of function for several uncharacterized cytochromes.

Reference

1. Beliaev, A. S., D. M. Klingeman, J. A. Klappenbach, L. Wu, M. F. Romine, J. M. Tiedje, K. H. Nealson, J. K. Fredrickson, and J. Zhou. 2005. "Global transcriptome analysis of *Shewanella oneidensis* MR-1 exposed to different terminal electron acceptors," *J. Bacteriol.* 187:7138-7145.

* Presenting author

74

Genome-Wide Transcriptional Responses to Metal Stresses in *Caulobacter crescentus*

Gary L. Andersen[1]* (GLAndersen@lbl.gov), Ping Hu[1], Eoin L. Brodie[1], Yohey Suzuki[3], and Harley H. McAdams[2]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]Stanford University School of Medicine, Stanford, CA; and [3]Research Center for Deep Geological Environments, Ibaraki, Japan

Effective bioremediation of metal contaminated sites requires knowledge of genetic pathways for resistance and biotransformation by component organisms within a microbial community. The bacterium *Caulobacter crescentus* and related stalk bacterial species are known for their distinctive ability to live in low nutrient environments, a characteristic of most heavy metal contaminated sites. *Caulobacter crescentus* is also a model organism for studying cell cycle regulation with well developed genetics. We have identified the pathways responding to heavy metal toxicity in *C. crescentus* to provide insights for possible application of *Caulobacter* to environmental restoration. Using a custom Affymetrix GeneChip array designed by the McAdams laboratory at Stanford University, analyses of genome wide transcriptional activities of *C. crescentus* cells post exposure to four heavy metals (chromium, cadmium, selenium and uranium) presented significant knowledge how *Caulobacter crescentus* activates different mechanisms in response to various metal stresses. Surprisingly, at the uranium concentration close to the highest observed at the NABIR Field Research Center (200 µM), *Caulobacter* growth rate was not significantly affected and it was not until a concentration of 1 mM uranium that *Caulobacter* growth slowed. Under the same conditions, growth of *E. coli* K-12 was completely stopped and the growth of *Pseudomonas putida* KT2440 was drastically reduced. To investigate the possible uranium resistance mechanism utilized by *Caulobacter crescentus* we performed transmission electron microscope (TEM) with energy-dispersive x-ray spectroscopy (EDX) analysis and demonstrated that *C. crescentus* did not form any uranium-containing phosphate granules intracellularly. However, TEM images of whole cells of *C. crescentus* revealed extracellular precipitates associated with the cells. EDX spectra from cells and extracellular precipitates showed that while uranium is almost absent within cells, extracellular precipitates contain high concentrations of uranium, phosphorus and calcium, suggesting that the extracellular precipitates are composed mainly of these elements. Based on the chemical composition, the precipitates are thought to be the uranyl phosphate mineral autunite, a major source of naturally occurring secondary uranium ore and is known to persist under oxidizing conditions on a geological time-scale. Transcriptional analysis did show a protein candidate, which may involved in the uranium precipitation process. Two two-component systems were identified to be specifically up-regulated in response to uranium stress. One pair of knockout mutants was studied to identify their possible targets. We also identified differentially expressed transcripts from antisense strand of a predicted gene responding specifically to metal stresses. Further studies may elucidate functions of these transcripts. The combination of whole genome transcriptional analysis, phenotypic and genetic studies and advanced imaging provided powerful insights into mechanisms of uranium resistance mechanisms by *Caulobacter crescentus*.

* Presenting author

75

Host Gene Expression Responses: Unique Identifiers of Exposure to Biothreat Agents

Rasha Hammamieh[1]* (rasha.hammamieh@na.amedd.army.mil), Steven Eker[2], Patrick Lincoln[2], and Marti Jett[1]

[1]Walter Reed Army Institute of Research, Silver Spring, MD and [2]SRI International, Inc., Menlo Park, CA

We have developed bioinformatic tools to facilitate mining of high throughput genomic and proteomic data. Today's fast-growing sphere of bioinformatics, where retrieving precise information on massive datasets can reveal in-depth understanding of systems biology. Using our program GeneCite, scientists can interconnect two input files via any of the three available Boolean operators at NCBI web domain. After completion of a given search, GeneCite provides a summary of result briefing total number of hits etc., and two output files. First file provides literature citation counts for each given search key, while the other file offers hyperlinks for each query connecting the appropriate result page of the data source. The other tool, PathwayScreen, takes a list of Gene ID numbers and outputs a file listing the pathways that those genes are in and a link to any appropriate resources, namely BioCarta.com. The SRI team has developed approaches for model building to identify unique gene patterns that can distinguish among biothreat pathogenic agents.

76

High Throughput Fermentation and Cell Culture Device

David Klein and Stephen Boyer* (sboyer@gener8.net)

Gener8 Inc., Mountain View, CA

The focus of our Phase II SBIR project is the creation of a high-throughput screening platform capable of delivering the essential controls of a stirred-vessel bioreactor (pH, dissolved oxygen, and temperature) in a small-scale, inexpensive, robust, easy-to-use, disposable format. The result of our work to date is an array of 24 x 10ml reactors in an SBS plate format. The capabilities of current MicroReactor systems will be reviewed and industrial as well as GTL research applications will be discussed. Industrial applications include clone and media screening as well as simple factorial design-of-experiments studies. GTL research applications include parametric studies of microbial physiology, cultivation of poorly characterized organisms; other applications might include controlled and reproducible sample generation for coarse studies of the microbial proteome, transcriptome, physiome, and metabolome.

* Presenting author

77

Genome-Wide Biochemical Characterization of Plant Acyl-CoA Dependent Acyltransferases

Chang-Jun Liu* (cliu@bnl.gov), and Xiao-Hong Yu

Brookhaven National Laboratory, Upton, NY

Acyl-CoA dependent acyltransferases catalyze the transfer of aliphatic and/or aromatic acyl moiety from CoA thioester donor to the nucleophile (OH- or NH-) of acceptor molecules. In plants enzymatic *O*- or *N*-acylation reactions are central to both primary and secondary metabolism, and are essential for plant growth and development, and plant environmental interactions. Particularly, acyl-CoA dependent *O*-acylation participates in plant cell-wall polysaccharide biosynthesis (Teleman et al., 2003), and the formation and deposition of heartwood-forming secondary metabolites in tree species, implicating significant biotechnological applications in genetic manipulation of lignocellulosic properties and carbon sequestration. Consistent with their multifaceted biological roles in plant metabolism, development and disease resistance, acyl-CoA dependent *O*-acyltransferases comprise a large and highly divergent protein family, known as BAHD superfamily (St-Pierre and De Luca, 2000). This family of enzymes consists of two conserved motifs, HXXXD and DFGWG in their primary sequences. Based on these sequence signatures, we employed tblastn algorithm searching poplar genomics sequences (http://genome.jgi.psf-org) and Tomato (SOL genomics network), *Medicago truncatula* (http://www.tigr.org/tigr-scripts/tgi/) EST databases and identified approximate 48 BAHD members in poplar genome, 48 Unigene in tomato unigene database, and 38 Tentative Consensus in *Medicago truncatula* EST database. Functional annotation of these large numbers of putative acyltransferase genes only based upon the sequence similarity with a few function known acyltransferases was difficult and extremely unreliable, due to the function diversity and primary sequence divergence. To unequivocally characterize the biochemical functions of putative acytransferases in genome-wide, we developed a high throughput biochemical assay procedure that consists of efficient Gateway cloning for expression vector construction, magnetic Ni-particles and microdialysis for rapid protein extraction and purification, 96 well-plate formatted *in vitro* assay, and single-well, on-line product detection and identification by High Performance Liquid Chromatography-Electron Spray Ionization Mass Spectrometry. Adapting the developed method and combining with structural homology modeling based functional prediction, we have characterized 5 novel acetyl-CoA and malonyl-CoA dependent, (iso)flavonoid and anthocyanin biosynthetic acyltransferases from *M. truncatula* EST clones. The details of biochemical characterization will be discussed.

* Presenting author

78

The MAGGIE Project: A Mass-Based Platform for Protein and Metabolite Characterization

Gary Siuzdak[1] (siuzdak@scripps.edu), Sunia Trauger[1], Francis E. Jenney Jr.[2], Steven M. Yannone[3], Daojing Wang[3], Nitin S. Baliga[4], Steven R. Holbrook[3], Chris H.Q. Ding[3], Angeli Menon[2], John A. Tainer[1,3], and Michael W.W. Adams[2]

[1]Scripps Research Institute, La Jolla, CA; [2]University of Georgia, Athens, GA; [3]Lawrence Berkeley National Laboratory, Berkeley, CA; and [4]Institute for Systems Biology, Seattle, WA

We have developed mass-based approaches for characterizing molecular complexes from *Pyrococcus furiosus* that include microLC ESI-MS/MS, microLC ESI-TOF, MALDIMS and DIOS-MS as well as a novel nonlinear software platform. These approaches initially examine proteins from the bacteria using multi-dimensional LC ESI-MS, followed by analysis of separated complex fractions via DIOS-MS, MALDI-MS and microLC ESIMS/MS. Intact proteins of the separated complexes are also examined using MALDIMS. Among the approaches being developed are perfluorinated surfactant-enhanced desorption/ionization on fluorinated silicon to provide greater protein coverage for PTM analysis. In addition, a new affinity approach based on the selective fluorous-fluorous interaction between fluorous tagged analytes and the fluorous silylated porous silicon (pSi) surface. By employing a simple washing procedure, a mixture containing target analytes deposited on the fluorous silylated pSi surface are selectively captured and enriched by affinity purification thereby facilitating its analysis. Metabolite data is simultaneously being generated with these analytical tools to investigate their role in these interactions as well as metabolite characterization through a newly created an online database of metabolite information (http://metlin.scripps.edu/). This research is largely synergistic with the efforts of Michael Adams and his ability to trap reversible and dynamic complexes enabling their purification. The platform along with XCMS software developments and initial results will be presented on the protein and metabolite characterization.

79

Organisation of Heavy Metal Resistance Genes in the Four Replicons of *Cupriavidus metallidurans* CH34

Sébastien Monchy[1], Nicolas Morin[1], Mohammed A. Benotmane[1], Sébastien van Aelst[1,2], Max Mergeay[1,2], Tatiana Vallaeys[1,3], Ruddy Wattiez[4], Safyih Taghavi[5], John Dunn[5], and **Daniel van der Lelie**[5]* (vdlelied@bnl.gov)

[1]Belgian Center for Nuclear Studies, SCK/CEN, Mol, Belgium; [2]Université Libre de Bruxelles, Belgium; [3]INRA, Jouyen-Josas, France; [4]Université Mons-Hainaut, Belgium; and [5]Brookhaven National Laboratory, Upton, NY

Cupriavidus metallidurans (formerly *Ralstonia*) belongs to the phylum β-Proteobacteria and includes various isolates of soil bacteria adapted to harsh industrial biotopes. The genome of strain CH34 has been sequenced (Joint Genome Institute) and contains four replicons: two large plasmids pMOL28 (171 Kb) and pMOL30 (234 kb), and two megareplicons – a chromosome (3.9 Mb) that is especially rich in biosynthetic genes, and a megaplasmid (2.6 Mb). Analysis of *C. metallidurans* on the genome level against existing databases and through phylogenetic approaches, the transcriptome level using

* Presenting author

quantitative PCR and microarrays analysis, and the proteome level (using 2-D gel electrophoresis and mass spectrometry) indicate a high number of heavy metal resistance or -detoxification genes in comparison to other sequenced bacteria. These genes seem to be mainly associated with the two plasmids and the smaller megareplicon of *C. metallidurans.*

Before the CH34 genome sequence became available, research focused on the plasmid-encoded *czc* (Cd, Zn, Co), *cnr* (Co, Ni), *chr* (Cr), *cop* (Cu), *pbr* (Pb) and *mer* (Hg) metal resistance operons, who had been identified via phenotypical analysis of plasmid-cured derivatives and cloning strategies. The whole genome sequence turned out to be a valuable resource for the identification of additional metal resistance systems, such as the *ars* operons located on the chromosome and several additional heavy metal efflux systems and their regulators.

In this communication we concentrate on the metal resistance determinants located on pMOL28 and pMOL30. Microarrays containing all ORFs as identified on the JGI draft sequence were hybridized with Cy3 and Cy5 labeled cDNA obtained from CH34 after induction with several heavy metals (30 min. induction with 0.5mM Cd, 5µM Hg, 0.8mM Zn, 0.4 mM Pb, 0.1mM Cu and 0.6mM Ni in 284 gluconate minimal medium at 30°C). We found 80 ORFs on a total of 161 located on pMOL28 and 134 ORFs on a total of 242 located on pMOL30 that were over-expressed in at least one metal condition.

Plasmid pMOL28 contains three clusters conferring resistance to nickel and cobalt (*cnr*), to chromate (*chr*) and to mercury (*mer* of Tn*4378*). These three clusters constitute a 35 kb region which is flanked by IS*1071* on the *mer* side and a deleted form of IS*1071* on the *cnr* side[1]. As expected, the *cnrYXHCBAT* cluster responded to Ni, but surprisingly also to Cu and Cd.

On plasmid pMOL30, heavy metal resistances are clustered opposite of the replication origin of the plasmid. Among other minor determinants, this region contains the *czcNICBADRSE* cluster, a mercury transposon (Tn*4380*), the *pbrTRABCD* cluster, the newly identified *silABC* operon that responds to Cu, and a large cluster of 19 genes comprising the Cu-resistance operon (*copVTMKNSRABCDIJGFLQHE*)[2]. Expression analysis revealed that the *czc* cluster was specifically induced by Zn and Cd. For the other metal resistance clusters, cross-responses inductions were observed to a much broader range of metals than expected from phenotypical analysis. Transcription of the *cop* cluster was not only induced by Cu but also by Zn, Cd and Ni, while transcription of the *pbr* cluster was not only induced by Pb but also Zn.

For both plasmids, the *mer* clusters' transcription responded to Hg but also to Zn, Cd and Pb, In addition, the microarray analysis allowed use to identify hypothetical ORFs (including several potential signal peptides) whose expression was highly induced in the presence of specific heavy metals. We hypothesize that these peptides could be involved in coordinating a general metal response by *C. metallidurans* CH34 against heavy metals.

Sequencing of the *C. metallidurans* genome has been carried out at the Joint Genome Institute (JGI) under the auspices of the U.S. Department of Energy's Office of Science, Biological and Environmental Research Program and the by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098 and Los Alamos National Laboratory under contract No. W-7405-ENG-36. The draft sequence of the genome is now available at the http://genome.jgi-psf.org/draft_microbes/ralme/ralme.home.html web site.

References

1. M. Mergeay, S. Monchy, T. Vallaeys, V. Auquier, A. Benotmane, P. Bertin, S. Taghavi, J. Dunn, D. van der Lelie, and R. Wattiez. (2003) "*Ralstonia metallidurans*, a bacterium specifically adapted to toxic metals: towards a catalogue of metal-responsive genes," *FEMS Microbiol Reviews*, 27: 385-410

* Presenting author

2. S. Monchy, M. A. Benotmane, R. Wattiez, S. Van Aelst, V. Auquier, B. Borremans, M. Mergeay, S. Taghavi, D van der Lelie, T. Vallaeys. (2005) "Transcriptomic and proteomic analyses of the pMOL30 encoded copper resistance in *Cupriavidus metallidurans* strain CH34," *Microbiology*, (Submitted)

80

Geobacter Project Subproject II: Expression of *Geobacteraceae* Genes Under Diverse Environmental Conditions

Dawn Holmes[1]* (dholmes@microbio.umass.edu), Brad Postier[1], Regina O'Neil[1], Kelly Nevin[1], Barbara Methe[2], Jessica Butler[1], Shelley Haveman[1], and **Derek Lovley**[1]

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

The purpose of this research is to describe genome-wide patterns of gene expression in *Geobacteraceae* species exposed to a variety of environmental conditions. These results are important not only for defining the metabolic state of *Geobacteraceae* in environments of interest but also provide the data needed to further elucidate gene function, determine regulatory networks, and further refine *in silico* models to predict the metabolism and growth of *Geobacteraceae* during *in situ* uranium bioremediation or harvesting electricity from organic wastes. Gene expression patterns from a diversity of pure cultures of *Geobacteraceae* grown under a wide range of environmentally relevant conditions was evaluated. This information was used to identify key genes which could aid in diagnosing the physiological state and rates of metabolism of *Geobacteraceae* in subsurface environments by quantifying *in situ* levels of transcripts for these genes.

Key to the inexpensive analysis of gene expression in multiple *Geobacteraceae* genomes was the development of methods to take advantage of the Combimatrix® microarray technology. With the Combimatrix® electrochemical synthesis method, arrays can be custom-made for individual experiments at lower cost than for other platforms. This affords the flexibility of readily evaluating gene expression in any pure culture for which a genome is available or custom-designing arrays for analysis of environmental transcripts, without the large investment in microarray synthesis required with other platforms. Gene expression studies conducted with the Combimatrix® arrays gave results comparable to several other platforms we had routinely used in the past with lower cost and higher flexibility.

Numerous microarray studies were conducted on pure cultures of *Geobacter sulfurreducens*, *Geobacter metallireducens* and *Pelobacter carbinolicus* grown under different conditions of electron donor or electron acceptor availability in order to elucidate genes involved in electron transfer to various electron acceptors or in the metabolism of electron donors. For example, with the recent availability of the complete genome sequence of *G. metallireducens* it was possible to conduct microarray studies under a variety of growth conditions that helped elucidate mechanisms for environmentally significant processes such as dissimilatory nitrate reduction to ammonia and the metabolism of aromatic compounds. Comparison of cultures of *P. carbinolicus* grown under fermentative versus Fe(III)-reducing conditions are helping to identify components in electron transfer to Fe(III) that are conserved among the *Geobacteraceae*. Studies in which chemostat cultures of *G. sulfurreducens* were provided with low levels of oxygen or limited by nutrient availability revealed genes diagnostic of oxidative stress or nutrient limitation. Comparison of electron donor versus electron acceptor limiting conditions have elucidated gene expression patterns diagnostic of *Geobacter* species becoming limiting for electron acceptor during *in situ* uranium bioremediation.

* Presenting author

Although chemostat cultures are able to provide physiologically consistent cells which are ideal for comparing gene expression under different growth conditions, chemostats are not the most accurate representation of the subsurface sediments in which *Geobacter* species grow during *in situ* uranium bioremediation or on the surface of energy-harvesting electrodes. A major advance has been the development of techniques for extracting mRNA and quantifying gene transcript levels from cultures grown in subsurface sediments or on electrodes. For example, *G. metallireducens* was grown in an artificial sediment of synthetic poorly crystalline Fe(III) oxide or in actual subsurface sediments from the *in situ* uranium bioremediation field site in Rifle, Colorado. In both instances there was significantly higher expression of genes associated with motility and chemotaxis in these cultures than in cultures grown on soluble electron acceptors. These results are significant because they provide insight into the unexpected result in field studies that many of the *Geobacter* involved in *in situ* uranium bioremediation are planktonic, rather than attached to the sediment particles, as was previously considered.

Down-regulated genes during growth in the sediments included several NADH dehydrogenase proteins, a number of ribosomal proteins, and other proteins involved in translation, transcription, and cell division. This reflects the slower growth of *G. metallireducens* under these more environmentally relevant conditions and reveals genes whose levels of expression might be used to estimate rates of growth of *Geobacter* species in the subsurface. During growth in sediments, genes for nitrogen fixation and phosphate uptake were more highly expressed, demonstrating that *in situ* uranium bioremediation might be limited by nutrient availability. This corresponds with measurements of transcript levels during *in situ* uranium bioremediation. Methods for overcoming these limitations can now be studied with genome-wide monitoring of gene expression in the sediment cultivation systems.

Analysis of gene expression during growth on electrodes was expanded from the studies reported last year to include a greater diversity of environmental conditions. As will be detailed in the presentation, the results from these studies have resulted in a model for the electrical contacts between the cell surface and electrodes.

These pure culture studies have provided significant guidance as to how best to monitor the *in situ* metabolic state of *Geobacteraceae* in environments of interest. As noted in our abstract on *Geobacteraceace* genomic sequences in these environments, the heterogeneity in genome sequences across time and environments requires that focus be placed on key, highly conserved genes that are representative of key types of metabolism. Our library of primers that can be used for quantitative RT-PCR analysis to track transcript levels of key genes in subsurface environments continues to build. Examples, of how measuring the transcript levels of genes diagnostic of limitation for nitrogen and phosphorous, oxidative stress, and rates of central metabolism have helped define the physiological state of *Geobacter* species during *in situ* uranium bioremediation at the field study site in Rifle, Colorado will be detailed.

81

Comprehensive Study of *recA* Expression in *Deinococcus radiodurans* with Single Cell and Population Level Analyses

Emily H. Turner* (emilyt@u.washington.edu), Haley Pugsley, and **Norman J. Dovichi**

University of Washington, Seattle, WA

The goal of this project is the development of instrumentation and methodologies for the analysis of biologically relevant proteins in single bacteria. We have built an instrument that performs capillary electrophoresis with laser-induced fluorescence, capable of detecting fluorescently-tagged proteins at the sub-zeptomole level (1 zeptomole = 600 molecules) in single eukaryotes and prokaryotes. The data obtained at the single cell level is complimented with population level analyses by capillary electrophoresis, flow cytometry and fluorescence microscopy. The combination of single cell and population level data provides a comprehensive analysis of protein expression and distribution across a population.

We are currently studying expression of the RecA protein in *Deinococcus radiodurans*. *D. radiodurans* displays an extraordinary capacity to repair DNA damage. Binding of the RecA protein to single-stranded DNA induces the SOS response to DNA damage; RecA has been identified as key to the survival of *Deinococcus* following high levels of DNA damage. The population-wide expression distribution of RecA, and the relationship of this distribution to DNA damage tolerance, is not well characterized. We have produced a novel recA/eGFP construct in *D. radiodurans*, strain MaHa01, to analyze *recA* expression with established and newly developed proteomic technologies.

Capillary electrophoresis with laser-induced fluorescence detection. Capillary electrophoresis is used to detect recA/eGFP expression in single *D. radiodurans* and across a population. In the single bacterium studies, a cell of interest is identified with fluorescence microscopy and injected into a small bore capillary (10 μm inner diameter), where lysis occurs. Separation of cellular components occurs with the application of high voltage, and results in rapid and ultrasensitive detection of released eGFP. This analysis resolves eGFP fluorescence from native autofluorescent components, and is capable of detecting eGFP expression in a single *Deinococcus* at the sub-zeptomole level.

Capillary electrophoresis is also used to characterize relative expression distributions of r*ecA* across a population. Intact cells are continuously injected into the capillary; as the bacteria exit the capillary and pass through the laser, intracellular eGFP fluorescence is detected. Changes in *recA* distribution across a population are rapidly determined after inducing DNA damage. Performing this analysis, which is analogous to flow cytometry, produces population-level data from the same instrument that measures our single cell data.

Flow cytometry. To induce DNA damage in MaHa01, cultures are exposed to mitomycin C (MMC). In future studies, UV radiation and bleomycin exposure will be used to induce DNA double-strand breaks. Flow cytometry is used to sort a MMC-exposed MaHa01 population into groups with high and low r*ecA* expression. After a 96 hour recovery period, these groups are re-exposed to MMC, and their r*ecA* expression is measured. An adaptive response is observed for both high and low *recA* groups. Preliminary data suggests the presence of two *recA* expression distributions within a population; further study of these distributions is planned.

Fluorescence microscopy. The growth of *D. radiodurans* in single, pair and tetrad groupings presents a challenge in quantifying protein expression across a population. Image analysis of fluorescence

* Presenting author

micrographs provides r*ecA* expression distributions for single, pair and tetrad MaHa01 populations. Using fluorescence microscopy, we have also observed stochastic r*ecA* expression between members of a tetrad.

The use of the recA/eGFP construct with these technologies provides comprehensive determination of the r*ecA* expression distribution in *D. radiodurans* and the effects of exposure to DNA damage on this distribution. We have recently developed a *D. radiodurans* construct that expresses eGFP under the control of the heat-shock *pGro* promoter. Similar analyses of *pGro* expression will be carried out in future work.

82

Directed Evolution of Radioresistance in a Radiosensitive Species

J.R. Battista[1]* (jbattis@lsu.edu) and Michael M. Cox[2]

[1]Louisiana State University and A & M College, Baton Rouge, LA and [2]University of Wisconsin, Madison, WI

Several bacterial species exhibit extraordinary resistance to ionizing radiation, surviving doses of 5,000Gy or higher without loss of viability. Although many hypotheses have been advanced to explain this radioresistance, very little is known about the specific mechanisms involved. To further our efforts to explain ionizing radiation resistance, we have taken the radiosensitive species, *Escherichia coli* strain MG1655, and subjected it to high dose ionizing radiation with the intent of generating a radioresistant strain that can be more easily studied than naturally radioresistant species. The protocol consisted of a series of selective steps in which successive exponential phase cultures were exposed to increasing doses of gamma radiation. The initial dose applied killed approximately 90% of the culture. The survivors were diluted into fresh growth medium and allowed to propagate. This process of irradiation and outgrowth was repeated for 21 generations. As the culture became more resistant to the effects of ionizing radiation the dose administered was increased. At the end of the study a purified resistant strain was recovered and its capacity to survive ionizing radiation evaluated. The evolved strain, which was designated 21-9 was approximately 500-fold more resistant at 5000Gy than its parent. The strain does not exhibit obvious phenotypic differences from MG1655, growing with normal kinetics in rich media at 37C. Analysis of genome restitution post-irradiation indicates the cells suffer DNA double strand breaks and those breaks are repaired with kinetics similar to those reported for irradiated cultures of *D. radiodurans*. 21-9 was re-sequenced using comparative genome sequencing, a microarray hybridization-based method developed by NimbleGen Systems Incorporated, to find mutations in the strain's genome. Sixty three differences were found in the genome of 21-9 relative to MG1655; 62 point mutations and one large deletion associated with the excision of the e14 prophage. Current efforts are focused on evaluating the role of these changes in ionizing radiation resistance.

* Presenting author

83

Development of a *Deinococcus radiodurans* Homologous Recombination System

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

J. Craig Venter Institute, Rockville, MD

A major goal of our Institute is to rationally design synthetic microorganisms that are capable of carrying out any required functions. One component of 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 severely 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 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, one approach is to assemble the homologous recombination activity by purifying and characterizing the analogous recombinant proteins. However, it is probable that not all genes that play a major role in homologous recombination have been identified by annotation. To overcome this potential obstacle, we are also 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. Progress made towards our goals will be presented.

84

Studies on the Fe Acquisition Mechanisms in *Nitrosomonas europaea* Derived from the Genome

Xueming Wei[1], Neeraja Vajrala[1], Norman Hommes[1]* (hommesn@onid.orst.edu), Cliff Unkefer[2], Luis Sayavedra-Soto[1], and **Daniel J. Arp**[1]

[1]Oregon State University, Corvallis, OR and [2]Los Alamos National Laboratory, Los Alamos, NM

Nitrosomonas europaea is a chemolithotrophic bacterium that can grow solely on ammonia (NH_3) and carbon dioxide (CO_2)[2]. The genome of *N. europaea* consists of ~2460 protein-encoding genes[1]. We have been utilizing the genome information to guide the studies on its unique Fe requirement and uptake systems. Fe is often a limiting factor for the growth of most bacteria because in aerobic environments, Fe exists predominantly in the insoluble ferric form (solubility in H_2O at pH 7.0 is 10^{-18} M).

The *N. europaea* genome reveals that up to 4% of the coding genes are dedicated to the transport of Fe, yet it noticeably lacks genes for siderophore biosynthesis[1]. There are 22 sets of genes that are

* Presenting author

organized similarly to the *fecI/fecR/fec*A system (genes for σ-factor/anti σ factor/TonB-dependent Fe-siderophore receptor/transducer), and 20 additional *fecA*-type genes that do not have associated *fecI/fecR* genes. All 22 genes in the first group encode outer membrane (OM) siderophore transducers that have an N-terminal extension, while 18 of the 20 genes of the second group code for OM siderophore receptors (lack of N-extension). These OM siderophore transducers/receptors are biochemically and phylogenetically diverse. 13 of the 42 OM receptor/transducer genes are either truncated or interrupted by IS elements and frame shifts and are likely not functional. Parallel to the large number of Fe acquisition-related genes, over 2% of the genome encodes proteins for heme and cytochrome biosynthesis and proteins with Fe-S centers. This high number of genes for Fe acquisition and Fe-containing proteins is consistent with the life style, especially the energy metabolism, of *N. europaea*. Our study has confirmed that *N. europaea* has higher contents of cellular Fe and cytochromes than common species such as *E. coli*.

We have determined the Fe requirement for *N. europaea* growth, and the effect of Fe limitation on cell physiology. Reverse transcriptase (RT)-PCR showed that 60% of the functional genes were expressed under either Fe-limited (0.2 µM) or Fe-replete (10 µM) conditions. The mRNA levels of a few genes appeared to be higher in Fe-replete cell than in Fe-limited cell. Four of the genes were expressed at much higher levels under Fe-limited condition than Fe-replete condition, all of which are genes encoding the siderophore receptors highly induced in Fe-limited cell that were identified by MS/MS. The expression of these genes at transcriptional level shows a diverse response to Fe availability. PAGE analysis also showed that several OM proteins were expressed at much higher levels under Fe limitation (0.2 µM Fe) than under Fe-replete (10 µM Fe) conditions. We have determined the identities of the differentially expressed OM proteins by HPLC tandem mass spectrometry (LC/MS/MS) analysis. Majority of these proteins are TonB-dependent receptors for siderophores such as ferrichrome and catechol-type siderophores. Included in these were proteins encoded by the four genes identified by RT-PCR as being highly expressed under Fe-limited condition. Interestingly, all of these OM proteins are true siderophore receptors that lack the N-terminal extension characteristic of the OM siderophore transducer family, and all are encoded by genes that do not have cognate σ-factor/ anti σ-factor genes. Three of the six genes encoding these receptors are the only ones of the 29 intact siderophores transducer/receptor genes that are preceded by a putative Fur box, suggesting the possibility of regulation by Fur (ferric uptake regulator). An OM porin OmpC, a multicopper oxidase, and a type II secretion pathway protein were also among the highly expressed proteins in Fe-limited cells. Both OM porin and multicopper oxidase could be involved in Fe uptake. These results provide evidence that under Fe deficient conditions, *N. europaea* up regulates the expression of certain Fe-acquisition-related proteins.

The addition of exogenous siderophores to Fe-limited medium increased *N. europaea* growth (total cell mass), suggesting its capability of using external siderophores for efficient Fe uptake. By LC/MS/MS analysis, we have also identified two OM transducers (encoded by NE1097/1088, *foxA* homologues) specific for the siderophore desferal, and they were expressed only in Fe-limited, desferal-containing medium, indicating that the expression required the induction by desferal. Both single and double mutants with disrupted desferal transducer genes have been created. Characterization of these mutants showed that the double mutants (with both genes inactivated) could not grow in Fe-limited, desferal-containing medium. Interestingly, the mutant with a disrupted gene NE1097 has the same phenotype as the double mutant, but single mutants with a defective gene NE1088 was able to grow in desferal-containing medium only when Fe level was raised to ~1.0 µM (5x[Fe] of Fe-limited medium). These results suggest that the acquisition of desferal-bound Fe needs functional desferal transducers, providing direct biochemical and genetic evidence for the functionality of the putative siderophore transducer genes in *N. europaea*. This result, together with the results from siderophore feeding experiments and elevated production of OM siderophore receptors under Fe limitation, re-enforces the notion that *N. europaea* can compete for Fe-loaded siderophores secreted by other microbes in its natural environments.

A putative regulatory mechanism for the desferal uptake system in *N. europaea* is proposed based on the genetic analysis and phenotypical behaviors of the desferal transducer mutants. The genome sequence shows that NE1097 exists with cognate *fecIR*-type genes, while there is no cognate *fecIR*-type gene preceding NE1088. Both OM desferal receptors are TonB-dependent transducers which can interact with anti σ-factor, which affects sigma factor, for the regulation of the expression of the systems. Based on these results and genetic information, it is likely that binding of Fe-loaded desferal to the transducers triggers the interaction between the transducers and the anti σ factor, and only the transducer encoded by NE1097 could interact with its cognate anti σ factor to turn on the expression of both genes. It also likely that the sigma factor encoded by the gene cognate to NE1097 could also interact with the promoter of gene NE1088 to turn on its expression.

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. 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.

85

Large Scale Genomic Analysis for Understanding Hydrogen Metabolism in *Chlamydomonas reinhardtii*

Michael Seibert[1]* (mike_seibert@nrel.gov), Florence Mus[2], Alexandra Dubini[1,3], Maria L. Ghirardi[1], Matthew C. Posewitz[3], and Arthur R. Grossman[2]

[1]National Renewable Energy Laboratory, Golden, CO; [2]Carnegie Institution of Washington, Stanford, CA; and [3]Colorado School of Mines, Golden, CO

While many taxonomically diverse microbes have the ability to produce H_2, only certain photosynthetic organisms, including the green alga, *Chlamydomonas reinhardtii*, are able to directly couple water oxidation to the photoproduction of H_2. A fundamental understanding of the metabolism in this prototype alga might enable the future development of a sustainable system for biological H_2 production. To work toward this understanding, we are exploiting whole genome sequence information and genomic tools that have been newly developed for *C. reinhardtii*. The completion of the *C. reinhardtii* genome sequence as part of the DOE Office of Science's Genomics:GTL Program facilitated the recent development of a high-density DNA microarray at the Carnegie Institution. This array is based on synthetic ~70 mers that represent approximately 10,000 unique *C. reinhardtii* genes, and is a powerful tool that can be used to thoroughly explore genome-wide changes in cellular transcript levels that accompany the acclimation of *C. reinhardtii* to conditions facilitating H_2 production.

Hydrogenase (the enzyme that catalyzes H_2 production) activity in *C. reinhardtii* is induced by anaerobiosis, achieved either in the dark by inert gas purging of O_2 or in the light by depriving cultures of sulfate. The latter attenuates photosynthetic O_2 evolution, which allows *C. reinhardtii* to metabolize any residual O_2 in the culture vessel. The resultant anaerobic environment sustains photoproduction of volumetric amounts of H_2 for 4 days in batch cultures. Under sulfur-deprived conditions, algal cultures

* Presenting author

exhibit a mixed metabolic state in which anaerobic fermentation, oxygenic photosynthesis and aerobic respiration co-occur. Our goal is to understand the underlying physiological processes that enable *C. reinhardtii* to sustain H_2 production, and toward this goal we are examining changes in transcript abundance and the establishment of protein networks that accompany the development of algal H_2 production. The use of high density DNA microarrays provides an initial view of the ways in which a cell may modulate its metabolism under different environmental conditions, and a 3,000 element array was recently used to examine sulfur- and phosphorus-deprivation responses in *C. reinhardtii*. The new 10,000 element array represents over half of the *C. reinhardtii* transcriptome and will provide more comprehensive information on ways in which this alga adjusts to conditions that sustain H_2 production.

Our initial studies are focused on characterizing differential gene expression in WT cultures that are aerobically grown and then anaerobically acclimated. Experimental protocols using qPCR have been established to rigorously quantify transcript levels for the *HydA1* and *HydA2* structural [FeFe]-hydrogenase genes and the *HydEF* and *HydG* [FeFe]-hydrogenase assembly genes in *C. reinhardtii* . Moreover, we have used qPCR to investigate relative transcript abundance of several genes involved in glycolysis and associated with the acclimation of the cells to sulfur-deprived growth conditions. Using *C. reinhardtii* strain CC425, the qPCR data indicate that there is a dramatic increase in abundance of the transcript encoding pyruvate ferrodoxin oxidoreductase (85-fold under appropriate conditions) during dark anaerobiosis, and that increased levels of transcripts for *HydA* structural genes (45-fold for *HydA1* and 25-fold for *HydA2*) and hydrogenase assembly genes (90-fold for *HydG* and 40-fold for *HydEF*) develop under anoxic conditions. Initial microarray data examining the differential expression of genes following dark-anaerobic induction were obtained. These data demonstrate significant changes in the transcript levels of several genes associated with signal transduction, transcriptional regulation, translational regulation, posttranslational modification, as well as photosynthesis, electron transport, proton transport, fermentation, stress response physiology, and a variety of other metabolic processes. Interestingly, the transcripts for several ribosomal proteins increase when the cells experience anaerobic conditions, indicating the possibility of significant changes in protein synthesis during anoxia. We also observe elevated levels of nitrate reductase transcript, which may reflect the establishment of a competing pathway for electrons away from H_2 production. Finally, transcripts from genes associated with oxidative stress also rise during anaerobiosis. These data provide the first insights into the metabolic pathways utilized by *C. reinhardtii* and the genome-wide changes in gene transcription that occur as this alga acclimates to an anoxic environment.

In addition to examining WT cultures, we have isolated several *C. reinhardtii* mutants at NREL (under another DOE Office of Science program) with attenuated H_2-photoproduction activity. We will compare gene expression profiles from these mutants with the WT under appropriate conditions. One such mutant lacks a functional *HydEF* gene, which is required to assemble an active hydrogenase enzyme. This *hydEF-1* mutant is the only reported *C. reinhardtii* strain that is unable to produce any H_2 at all. Since *hydEF-1* is specifically disrupted in its ability to synthesize an active hydrogenase, any gene that is differentially expressed in this mutant should be a consequence of the mutant's inability to photoproduce H_2. Another *C. reinhardtii* mutant, *sta7-10*, is unable to accumulate intracellular starch. Interestingly, this mutant shows aberrant induction of hydrogenase transcript accumulation and attenuated H_2-photoproduction activity during anaerobiosis.

In sum, we have begun to develop a global understanding of factors that promote H_2 production during anaerobiosis by analyzing transcript profiles from WT cultures of *C. reinhardtii*. This work is elucidating the biochemical pathways utilized by *C. reinhardtii* during anaerobiosis and will provide insights into how mutants, altered in normal H_2 metabolism, acclimate to anaerobiosis. More detailed knowledge of the metabolic and regulatory context that facilitates H_2 production will be necessary to understand and ultimately correct current limitations in H_2-production yields.

86

Single-Molecule Imaging of Macromolecular Dynamics in a Cell

Jamie H.D. Cate[1,2] and Haw Yang[1,2]* (hawyang@berkeley.edu)

[1]Lawrence Berkeley National Laboratory, Berkeley, CA and [2]University of California, Berkeley, CA

We are taking several approaches to make the tools that will be needed for single-molecule imaging of macromolecule dynamics in living cells. An apparatus that tracks a single moving nanoparticle in 3D while providing concurrent sequential spectroscopic measurements has been developed. The design is based on confocal microscopy and is the first step towards correlating the reactivity of a single molecule with its spatial location in cells. One critical element in tracking single molecules in cells is the optical probes that contrast the molecule against cellular background. This goal is approached by luminescence engineering of semi-conducting quantum dots (Qdot). We report detailed characterizations of such engineered Qdots at the single particle level. The experimentally obtained lifetime-intensity correlation maps suggest that Qdot charging states are continuously distributed, and provide the physical foundation for luminescence engineering by synthesis. To place exogenous probes inside a bacterial cell in a controlled way, reliable methods to overcome the membrane barrier have to be developed. We characterize the two most commonly used methods, electroporation and heat shock, at the single-cell level. It was found that probes introduced via electroporation enter a cell primarily through the pole whereas those introduced via heat shock through the newly synthesized membrane. These observations provide important clues for controlled placement of exogenous probes into bacterial cells. Finally, we report the successful construction of a microfluidic mixer that allows studies of biological macromolecules under crowding conditions that are comparable to those inside a cell. This mixer platform is critical in validating any models that may form from our future *in cellular* studies. With the high-resolution ribosome structure solved, the new tools we are developing will form the basis for *in cellular* studies of protein synthesis machinery.

87

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

Gang Bao[1,2]* (gang.bao@bme.gatech.edu), Grant Jensen[3], Shuming Nie[1,2], and Phil LeDuc[4]

[1]Georgia Institute of Technology, Atlanta, GA; [2]Emory University, Atlanta, GA; [3]California Institute of Technology, Pasadena; and [4]Carnegie Mellon University, Pittsburgh, PA

We have been developing quantum-dot (QD) based strategies for imaging and identification of individual proteins and protein complexes in microbial cells. Currently, there is a lack of novel labeling reagents for visualizing and tracking the assembly and disassembly of multi-protein molecular machines. There is no existing method to study simultaneous co-localization and dynamics of different intra-cellular processes with high spatial resolution. As shown in Figure 1, the multifunctional quantum-dot bioconjugates we develop consisting of a quantum dot of 2-6 nm in size encapsulated in a phospholipid micelle, with delivery peptides and protein targeting ligands (adaptors) conjugated to the surface of the QD through a biocompatible polymer. After internalization into microbial cells, the adaptor molecules on the surface of QD bioconjugates bind to specific target proteins or protein complexes that are genetically tagged. Optical imaging is used to visualize the localization, trafficking

* Presenting author

and interaction of the proteins, resulting in a dynamic picture but with a limited spatial resolution (~200 nm). The same cells is 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 are tested to optimize the specificity and signal-to-noise ratios of protein detection and localization. This innovative molecular imaging approach integrates peptide-based cellular delivery, protein targeting/tagging, light microscopy and electron microscopy.

To achieve the goals of this DOE GTL project, we have successfully synthesized core-shell and alloyed CdHgTe quantum dots (QDs) for dual-modality optical and EM imaging. This new class of QDs contains Hg, a heavy element that is often used in x-ray and electron scattering experiments, allowing studies of cellular structures at nanometer resolution. We have also linked QDs to a chelating compound (nickel-nitrilotriacetic acid or Ni-NTA) that quantitatively binds to hexahistidine-tagged biomolecules with controlled molar ratio and molecular orientation.

We have tested a number of methods for delivery of QD probes into living cells, and identified the advantages and limitations of each method. For example, we explored the possibility of delivering QD probes into yeast and *E. coli* with high efficiency using different methods, including peptide-based delivery, heat shock, and the use of anti-microbial/permeabilizing agents. Specifically, we performed a preliminary study of peptide-based delivery of QD bioconjugates into yeast and *E-coli* using three different peptides, TAT, polyArg, and a peptide (*ArgSerAsnAsnProPheArgAlaArg*) that has been used for delivering GFP into yeast *S. cerevisiae*. We have also tested different tagging strategies including tetracysteine/FIAsH, SNAP tag, Histidine/Ni-NTA and Histidine-peptide.

As part of our effort to develop QD-based technologies to identify and track individual protein complexes in microbial cells, we have performed preliminary optical imaging studies of single QDs delivered into living cells. Using a spinning-disk confocal microscope, we have succeeded in imaging single QD probes delivered into the cytoplasm of living cells. Several lines of evidence support that the QDs in cells are indeed single: (a) these QDs have similar brightness and spot size; (b) the brightness of these QDs is not higher than that of single QDs on a coverslip; and (c) the intracellular QDs show intermittent on/off light emission (called blinking), a characteristic of single dot behavior. We have also developed computation algorithms for two-color colocalization and correlation tracking of QD probes. As an alternative, we successfully imaged individual 10 nm gold nanoparticles and established the darkfield optical imaging capability for cellular studies.

We are advancing electron tomography as a promising new tool to image protein complexes both *in-vitro* and *in-vivo* within small microbial cells. A new helium-cooled, 300kV, FEG, "G2 Polara" FEI

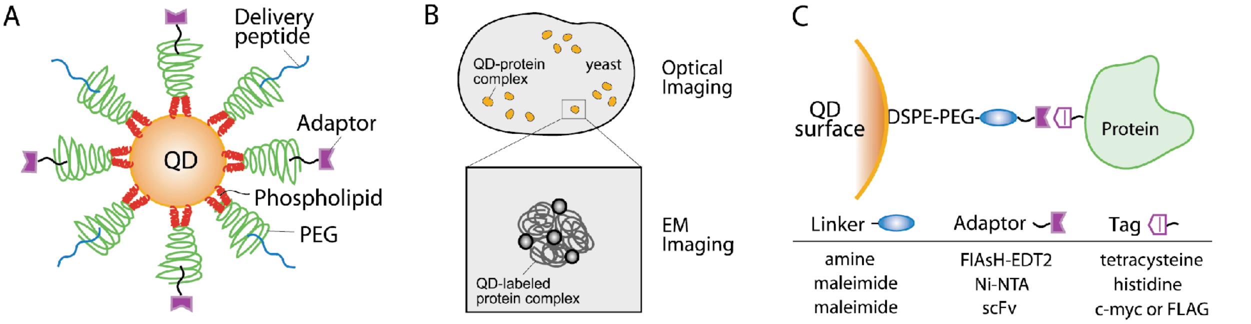


Figure 1. (A) Schematic illustration of a multifunctional quantum dot bioconjugate consisting of encapsulated QD with targeting adaptor and delivery peptide on its surface; (B) correlated optical and EM imaging of the same cell gives both temporal and spatial information on a protein complex; (C) possible conjugation and tagging strategies for optimizing detection specificity and sensitivity. Note that molecules are not drawn to the exact scale.

TEM at Caltech was used to image purified protein complexes, viruses, and whole bacterial cells. We pioneered the use of a new "flip-flop" cryorotation stage that allows dual-axis cryotomography, and developed a simple Perl-based system for distributed computation to handle the massive image processing demands that arise from imaging intact bacteria in 3D. These technological advances have allowed us to visualize directly cytoskeletal elements within small microbial cells and the domain structure of purified multienzyme complexes, both are key imaging goals of the genomes to life program. For example, we produced three-dimensional reconstructions of several different types of bacteria, including some of great interest to the DOE (*Magnetospirillum magneticum*, *Mycoplasma pneumonia*, and *Caulobacter crescentus*), with unprecedented resolution and authenticity. We imaged chemoreceptor clusters, flagella, pili, polyribosomes, and other ultrastructural details, and identified five unique patterns of cytoskeletal filaments bundles likely involved in cell shape determination, establishment of polarity, and chromosome segregation in *C. crescentus*.

As a model system to study protein localization, we have been investigating the migration of *Dictyostelium discoideum* under defined extracellular stimuli. We have utilized custom-fabricated microfluidic devices to stimulate a cell in local domains both with 2D and 3D control while simultaneously visualizing its response with fluorescent microscopy using quantum dots. We targeted quantum dots to CRAC and G-actin and analyzed them for co-localization of the GFP and quantum dot signals in *Dictyostelium*. This technique will be further combined with high-resolution electron microscopy imaging to visualize individual proteins and protein complexes.

Acknowledgement: This research is funded by a grant from DOE (DE-FG02-04ER63785).

88

Hyperspectral Imaging of Photosynthetic Pigment Molecules in Living Cyanobacterial Cells

Jerilyn A. Timlin[1], Michael B. Sinclair[1], Howland D.T. Jones[1], Linda T. Nieman[1], Sawsan Hamad[2], **David M. Haaland**[1]* (dmhaala@sandia.gov), and Wim F.J. Vermaas[2]

[1]Sandia National Laboratories, Albuquerque, NM and [2]Arizona State University, Tempe, AZ

Hyperspectral confocal imaging has the potential to revolutionize the quality and quantity of energy transfer information derived in situ from living, photosynthetic systems. We have developed this technology at Sandia, and applied it to intact, living cyanobacterial cells to follow energy transfer within the light-harvesting antenna of the photosynthetic apparatus. Light is absorbed by a number of different pigments in the cell, including phycobilins, chlorophyll *a*, and carotenoids. Phycobilins bound to specific proteins serve as an antenna complex (the phycobilisome) to funnel excitation energy to lower-energy pigments and eventually to a chlorophyll *a* molecule in a special protein environment, the photosynthetic reaction center. Phycobilin pigments and chlorophylls in excited states will usually transfer their energy to the reaction center chlorophyll, but can also decay by fluorescence emission or internal conversion. The fluorescence is in principle a signature of the pigments in the excited state, but the emission wavelengths of the various pigments are very close together.

Using hyperspectral fluorescence microscopy of intact cyanobacterial cells and multivariate analysis technology, we have identified five different fluorescent spectral signatures with maxima between 640 and 700 nm and mapped their location within the living cell (see Figure 1 for an example using

* Presenting author

wild-type cells). This unique imaging system captures a full emission spectrum at each image pixel and uses multivariate curve resolution (MCR) technology[1,2] to deduce the fluorescence spectrum and location of individual pigment species with single molecule sensitivity in three dimensions with a spatial resolution of 0.24 µm in X and Y directions, and 0.6 µm in the Z direction[3]. A single cell can be imaged in all three dimensions with 50 ms temporal resolution.

Upon excitation at 488 nm, six fluorescence components are resolved from the hyperspectral images of intact cells: three phycobilins (phycocyanin (PC), allophycocyanin (APC), and allophycocyanin-B (APC-B)), chlorophyll *a*, a weaker long-wavelength chlorophyll *a* component peaking at ~698 nm and an extremely weak short-wavelength component. The latter has resonance-Raman bands detectable with our imaging system that correspond to carotenoids, primarily β-carotene. These six spectral signatures are common to the wild type and mutant strains, but differences in their relative intensities and spatial locations between mutant strains make them useful for developing insight as to molecular level localization within the cell. For example, the 698 nm component is due primarily to photosystem I (PS I)-related chlorophyll (it is greatly decreased in mutants lacking PS I or strains depleted in chlorophyll) and is distributed much more evenly through the cell than the phycobilin components, which are concentrated around the periphery. By comparing spectra and distribution of components from the wild type and strains lacking PS I and/or ChlL, an enzyme needed for light-independent conversion between biosynthetic precursors of chlorophyll, individual fluorescence components can be assigned to specific emitting pigments.

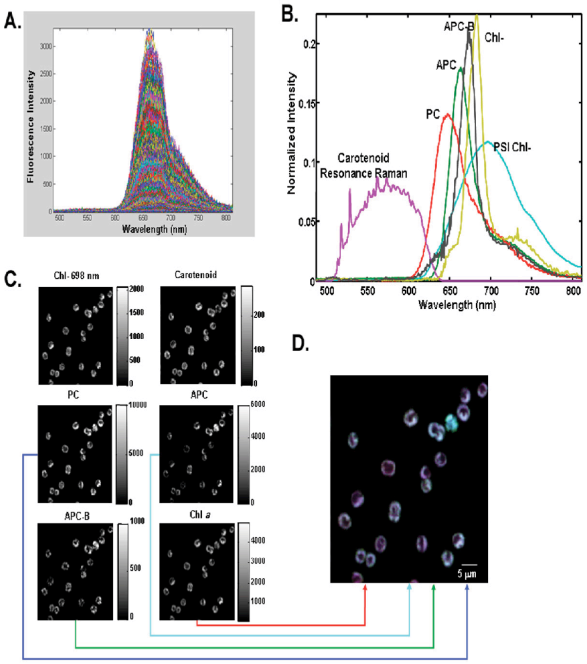


Figure 1: Information extraction from hyperspectral confocal images of live wild type *Synechocystis* sp. PCC 6803 cells. A. ~43,000 raw spectra from a hyperspectral data cube (spectra are from a 25µm x 25µm 2D slice taken from the larger 25µm x 25µm x 6µm 3D image) B. Multivariate curve resolution (MCR)-extracted pure spectral signatures from the hyperspectral data in A. C. Corresponding concentration maps indicating the spatial location of each of the components in B. D. Composite RGB image of phycocyanin (false colored blue), allophycocyanin (false colored cyan), allophycocyanin B (false colored green), and chlorophyll (685 nm peak, false colored red) in wild type *Synechocystis* sp. PCC 6803 cells.

The results obtained indicate a preferential localization of fluorescing phycobilin components around the periphery of the cells, whereas chlorophyll emission is more evenly distributed within the cell. Based on these and other results, our current explanation is that unattached, highly fluorescent phycobilisomes (*i.e.*, phycobilisomes that do not transfer energy to photosynthetic reaction center complexes) are toward the periphery of the concentric stack of thylakoids in the cell, whereas photosynthetic reaction center complexes are more evenly distributed among thylakoids.

We show here that the hyperspectral confocal imaging approach provides highly detailed information regarding sub-cellular localization of pigments in living cells with unparalleled resolution. We anticipate to be able to further develop and refine the

* Presenting author

technique and deconvolutions to be able to visualize and follow a large number of fluorescing cellular components in the cell.

References

1. P. G. Kotula, M. R. Keenan, J. R. Michael, *Microscopy & Microanalysis* **9**, 1 (2003).
2. J. R. Schoonover, R. Marx, S. L. Zhang, *Applied Spectroscopy* **57**, 154A (2003).
3. M. B. Sinclair, D. M. Haaland, J. A. Timlin, H. D. T. Jones, *Applied Optics*, submitted (2005).

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

89

The Cyanobacterium *Synechocystis* sp. PCC 6803: Membrane Biogenesis, Structure and Function

Wim Vermaas* (wim@asu.edu), Robert W. Roberson, Sawsan Hamad, Bing Wang, Zhi Cai, and Allison van de Meene

Arizona State University, Tempe, AZ

Cyanobacteria play a key role in global carbon fixation and energy conversion, and selected strains lend themselves very well to metabolic engineering and synthetic biology. One strain that is particularly useful in this respect is the unicellular cyanobacterium *Synechocystis* sp. PCC 6803 that is readily amenable to detailed structural and functional investigations as it has a known genome sequence, is easily transformed, can grow under a wide variety of environmental conditions, and has developed into the premier cyanobacterial model system for photosynthesis and respiration studies. A major advantage of *Synechocystis* is its ready availability of informative mutants with knock-outs of one or more (up to seven) genes and/or with overexpression of introduced genes. As *Synechocystis* was the first cyanobacterium to be sequenced, another major advantage of this organism is the abundance of bioinformatics resources geared toward this cyanobacterium (e.g., CyanoSeed developed with Genomics:GTL funding (http://theseed.uchicago.edu/FIG/organisms.cgi?show=cyano) and CyanoBase, http://www.kazusa.or.jp/cyanobase/Synechocystis/index.html). In our GTL-supported studies we focus on cell structure and cell physiology in *Synechocystis*, with particular emphasis on thylakoid membrane formation and on metabolism related to photosynthesis and respiration. New results on (a) thylakoid membrane biogenesis, (b) fluxes through central carbon utilization pathways, and (c) distribution mechanisms between carbon storage compounds will be presented in subsequent paragraphs. Together, these results help pave the way for metabolic engineering efforts resulting in improved bioenergy production and carbon sequestration.

The internal thylakoid membrane system of *Synechocystis* comprises about 80% of the total membrane content of the cell, and contains membrane protein complexes involved in both photosynthesis and respiration. Thylakoid organization is rather sophisticated, with membranes occurring in multiple layers along the periphery of the cell while often connected to a rod-like structure, the thylakoid center[1]. Thylakoid formation seems to be critically correlated with the presence of chlorophyll and not with the presence of photosynthetic complexes, as in a mutant where chlorophyll synthesis is under strict light control thylakoids are essentially absent after prolonged growth in virtual darkness, whereas thylakoid membranes form rapidly upon exposure to light. In contrast, mutants lacking both photosystems retain a significant amount of thylakoids. The molecular mechanism

* Presenting author

of thylakoid formation remains largely unknown, but some proteins that may be involved with thylakoid generation have been identified. Upon overexpression of one of these proteins we found more and closer spaced thylakoids protein in *Synechocystis*, along with novel membrane structures that may be instructive in understanding membrane biogenesis. These results indicate that we now are able to generate cyanobacterial strains with increased levels of the photosynthetic apparatus and thylakoids (useful biomass) per cell.

Metabolic engineering is aided by detailed insight into fluxes through main metabolic pathways. Thus far, flux data regarding central carbon utilization pathways are derived largely from a rather indirect approach of isotope labeling and monitoring the isotopic composition of end products such as amino acids. However, cyanobacterial carbon utilization is very complex, with sugar utilization by the pentose phosphate pathway and glycolysis immediately connected with carbon fixation pathways through the Calvin-Benson-Bassham cycle. Most steps in the pathways are reversible, and several steps involve the reorganization of C-C bonds, causing rapid isotope scrambling when providing uniformly labeled ^{13}C-glucose. With improved LC/MS (liquid chromatography/mass spectrometry) techniques we have now been successful in monitoring the mass redistribution of several sugar phosphates as a function of time after adding labeled glucose. Comparing these results under different growth conditions and in specific mutants lacking particular steps in the pathways yields a detailed insight regarding *in vivo* rates of key metabolic reactions in carbon utilization.

Synechocystis sp. PCC 6803 has two main carbon storage compounds: glycogen and polyhydroxybutyrate (PHB). The compound to be accumulated has been found to primarily depend on the environmental conditions. Glycogen is found under many conditions, but under –for example- nitrate limitation PHB can make up 5-10% of the dry weight of the cell. We have explored the metabolic reasons for this apparent dichotomy in the preferred carbon storage compound in *Synechocystis*. Based on results of comparative PHB accumulation analysis as a function of environmental conditions in different mutant strains, the level of reduction of specific redox carriers in the cell is found to be a key determinant for PHB accumulation. With this knowledge PHB production in *Synechocystis* can now be optimized.

These aspects of the project build on an excellent foundation of genomic and functional data regarding *Synechocystis* sp. PCC 6803, and together provide a solid basis for metabolic engineering of this cyanobacterium to enhance solar-powered carbon sequestration and bioenergy conversion.

Reference

1. van de Meene, A.M.L., M.F. Hohmann-Marriott, W.F.J. Vermaas, and R.W. Roberson (2006) "The three-dimensional structure of the cyanobacterium *Synechocystis* sp. PCC 6803," *Arch. Microbiol.* 184: 259-270.

Section 2

Metabolic Network Experimentation and Modeling

90 MEWG

Metabolic Engineering of Light and Dark Biochemical Pathways in Wild-Type and Mutant *Synechocystis* PCC 6803 Strains for Maximal, 24-Hour Production of Hydrogen Gas

P.S. Schrader[1], E.H. Burrows[2], F.W.R. Chaplen[2], and **R.L. Ely**[2]* (ely@engr.orst.edu)

[1]Yale University, New Haven, CT and [2]Oregon State University, Corvallis, OR

Photobiological production of H_2 from water has great appeal as an environmentally sustainable, long-term means to meet large, projected increases in demand, to provide energy and economic security for the U.S. and other nations, and to relieve environmental stresses related to fossil fuel use

This project is using the cyanobacterial species *Synechocystis* PCC 6803 as a model to pursue two parallel lines of inquiry initially, with each line addressing one of the two main factors affecting H_2 production in PCC 6803: NADPH availability and O_2 sensitivity. H_2 production in PCC 6803 requires a very high NADPH:NADP$^+$ ratio, that is, that the NADP pool be highly reduced, which can be problematic because several metabolic pathways potentially can act to raise or lower NADPH levels. Also, though the [NiFe]-H_2ase in PCC 6803 is constitutively expressed, it is reversibly inactivated at very low O_2 concentrations, reportedly due to binding of O_2 to the active-site. Largely because of this O_2 sensitivity and the requirement for high NADPH levels, a major portion of overall H_2 production occurs under anoxic conditions in the dark, supported by breakdown of glycogen or other organic substrates accumulated during photosynthesis. Also, other factors, such as N or S limitation, pH changes, presence of other substances, or deletion of particular respiratory components, can affect light or dark H_2 production. Therefore, in the first line of inquiry, under a number of culture conditions with wild-type (WT) PCC 6803 cells and a mutant with impaired type I NADPH-dehydrogenase (NDH-1) function, we are using H_2 production profiling and metabolic flux analysis, with and without specific inhibitors, to examine systematically the pathways involved in light and dark H_2 production. Results from this work will provide rational bases for metabolic engineering to maximize photobiological H_2 production on a 24-hour basis. In the second line of inquiry, we are using site-directed and random mutagenesis to create mutants with H_2ase enzymes exhibiting greater O_2 tolerance (and perhaps higher H_2 production activity). The objectives of the research are addressed via the following four tasks:

1. Evaluate the effects of various culture conditions (N, S, or P limitation; light/dark; pH; exogenous organic carbon) on H_2 production profiles of WT cells and an NDH-1 mutant;
2. Conduct metabolic flux analyses for enhanced H_2 production profiles using selected culture conditions and inhibitors of specific pathways in WT cells and an NDH-1 mutant;
3. Create PCC 6803 mutant strains with modified H_2ases exhibiting increased O_2 tolerance and greater H_2 production;
4. Integrate enhanced H_2ase mutants and culture and metabolic factor studies to maximize 24-hour H_2 production.

Task 3 is being conducted in parallel with Tasks 1 and 2; Task 4 will reflect the convergence of research performed in Tasks 1-3.

* Presenting author

91

Geobacter Project Subproject V: *In Silico* Modeling of the Growth and Physiological Responses of *Geobacteraceae* in Complex Environments

Radhakrishnan Mahadevan[1]* (rmahadevan@genomatica.com), Maddalena Coppi[2], Daniel Segura[2], Steve Van Dien[1], Bernhard Palsson[1], Christophe Schilling[1], Abraham Esteve-Nunez[2], Mounir Izallalen[2], and **Derek Lovley**[2]

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

The ultimate goal of the *Geobacter* Project is to develop genome-based *in silico* models that can be used not only to interpret environmental gene expression data in environments in which *Geobacteraceae* predominate, but also to predict with *in silico* studies, the outcome of various potential manipulations that might be made to optimize processes, such as *in situ* uranium bioremediation and harvesting electricity from waste organic matter, prior to conducting labor-intensive and expensive field experiments. Although the recently published *in silico* model of *Geobacter sulfurreducens* has been effective in providing explanations for important physiological phenomena and in guiding functional genomic studies, further development is necessary to improve predictions and make them more applicable to other *Geobacter* species.

For example, acetate is the key electron donor for *Geobacter* species during *in situ* uranium bioremediation and in the conversion of organic matter to electricity. Further analysis of the *in silico* metabolic model for *G. sulfurreducens* identified redundant pathways for acetate metabolism. . There are two acetate activation pathways encoded in the genome, the acetate kinase/phosphate transacetylase (Ack/Pta) pathway and the acetyl-CoA transferase (Ato), which plays a dual role in acetate activation and the TCA cycle. There are also, two enzymes catalyzing the synthesis of oxaloacetate, the TCA cycle enzyme, malate dehydrogenase (Mdh) and pyruvate carboxylase (PC), which catalyzes the conversion of pyruvate to oxaloacetate. Three reactions are present for the synthesis of acetyl-CoA from pyruvate: pyruvate dehydrogenase, pyruvate formate lyase and pyruvate ferredoxin oxidoreductase (Por) and three are possible pathways for synthesis of PEP involving pyruvate phosphate dikinase (PpdK), PEP synthase (PpS), and PEP carboxykinase (PpcK).

To evaluate the role of these pathways, five knockout mutant strains lacking elements of the various redundant pathways (Ato, Pta, Por, Mdh, PpcK) were constructed and evaluated along with the wild type for their ability to grow under twelve distinct environmental conditions (72 combinations) and the model predictions were compared to the results of the phenotypic analysis. The model predicted that *G. sulfurreducens* would be able to compensate for the absence of Ato by increasing flux through the Ack/Pta pathway and succinyl-CoA synthetase. However, failure of the Ato-knockout mutant to grow on acetate suggested that the succinyl-CoA synthetase was inactive. Similar constraints on metabolism were derived from the comparison of the *in vivo* phenotypes with the model predictions. Following the incorporation of these new constraints, the *in silico* model now correctly predicts the experimental result in 89% of the possible conditions providing highly accurate characterization of central metabolism in *G. sulfurreducens.*

Despite the augmentation of the constraints using the genetic data described above, the relative flux through the multiple PEP synthesizing pathways could not be resolved by the *in silico* model. In order to further characterize central metabolism and validate the model, carbon isotope (^{13}C) labeling studies were designed using the genome-scale *in silico* model and carried out initially in batch cultures of *G. sulfurreducens*. The *in silico* model was used to determine the optimal acetate labeling strategy for distinguishing the flux through the various phosphoenolpyruvate synthesizing pathways. The

* Presenting author

predicted optimal labeling ratio (70% unlabeled and 30% labeled acetate) was utilized for preliminary studies in batch culture and the flux through key reactions in central metabolism was calculated based on the distribution of the label in the amino acids. Further comparison of the experimentally measured flux distribution with model prediction under steady state conditions in a variety of growth media will aid in validating and refining the model.

The *in silico* model has also been used to predict the metabolic adaptation to several environmental perturbations including nitrogen limitation, electron acceptor limitation, and phosphate limitation. These predictions are currently being compared to experimental data, including physiological measurements and global gene expression data, with the goal of improving our understanding of the metabolism of *Geobacteraceae* under non-optimal but environmentally relevant conditions. For example, *in silico* modeling of growth under electron acceptor limiting conditions suggested that either formate or hydrogen production might serve as electron sinks. Although, formate was not detected in the media in response to acceptor limitation, there was increased hydrogen production. Another metabolic adaptation to this condition included an increase in acetate flux through the TCA cycle. Comparison of global gene expression data with model predictions of flux distribution revealed some evidence of increased TCA cycle flux, including upregulation of isocitrate dehydrogenase and downregulation of the phosphotransacetylase. However, there were several exceptions (downregulated acetate transporter, citrate synthase) indicating that flux changes may not be well correlated with transcription profiling data.

Another subject of current investigation is the incorporation of regulatory constraints in the metabolic model, as recent studies in *Escherichia coli* have shown that the addition of such constraints can improve the quality of predictions. The analysis of the existing gene expression data for *G. sulfurreducens* revealed putative regulatory interactions involved in heat shock response, and sulfate metabolism. These initial regulatory interactions will be integrated with the *in silico* metabolic model to design experiments that will optimally perturb the regulatory network. The integrated model will be used to pick the most informative environment changes and the transcription factors for deletion. The gene expression data obtained from these environments will be used to assemble the regulatory network and this process will be repeated to obtain a refined and integrated regulatory and metabolic network model. However, further investigation will be required to extend the results from the *G. sulfurreducens* model to characterize metabolism in other members of *Geobacteraceae.*

Although, recent comparative studies of *Geobacteraceae* genomes have indicated that the electron transport chain components are not fully conserved across the different members of this family, the majority of the elements of central metabolism are conserved among the various family members. Hence, the development of models of metabolism for other *Geobacteraceae* can be accelerated by leveraging the existing curated *G. sulfurreducens* model. An automated modeling pipeline has been redesigned from earlier prototypes to reconstruct the metabolic network for a new organism through a comparison of its genome with the organisms for which a high quality model is available. This has enabled the rapid construction of draft genome-scale models that can be manually curated further to obtain a complete model. We have utilized this pipeline to construct models of other *Geobacteraceae* including *G. metallireducens*, and *Pelobacter carbinolicus* based on the *G. sulfurreducens* model. The initial model of *G. metallireducens* contains about 566 genes, 514 reactions, whereas *P. carbinolicus* model contains 444 genes and 527 reactions. These models are currently being manually curated to ensure that the model can synthesize all essential biomass components. The construction of a comprehensive and physiologically validated *in silico* models of these and other organisms will create a database of metabolic functions that will be valuable for predicting the metabolic capabilities of environmental isolates and optimizing strategies for bioremediation.

It is expected that *Geobacter* species might readily be genetically modified to improve electricity generation because there has been no previous evolutionary pressure to select for this property. Model-

* Presenting author

based analysis suggested that the respiration rate and subsequently, the rate of electron transfer to electrodes could be increased with the introduction of an energy draining futile cycle. An additional ATP consuming reaction was added to *G. sulfurreducens* by introducing the gene for the cytosolic portion of the ATP synthase under the control of an IPTG-inducible promoter. When IPTG was added to culture media, cells had higher respiration rates and the current generation doubled. Electricity production might also be enhanced if *Geobacter* species could utilize more electron-dense fuels. Model simulations indicated that the addition of glycerol transport capability to *G. sulfurreducens* would enable glycerol utilization and this prediction was experimentally confirmed. These studies demonstrate that genome-based *in silico* modeling of microbial physiology can significantly aid in experimental design for strain improvement for practical applications.

92 MEWG

Development of Computational Tools for Analyzing and Redesigning Biological Networks

Priti Pharkya[1], Madhukar Dasika[1], Vinay Satish Kumar[1], Narayanan Veeraghavan[1], Patrick Suthers[1], Anthony Burgard[2], and **Costas D. Maranas**[1]* (costas@psu.edu)

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

The incredible growth in recent years of biological data at all levels has provided a major impetus for developing sophisticated computational approaches for unraveling the underlying complex web of protein, DNA and metabolite interactions that govern the response of cellular systems to intracellular and environmental stimuli. Even partial knowledge of these interconnections and interactions can facilitate the targeted redesign of these systems in response to an overproduction objective. In this poster, we will highlight our progress towards the development of computational frameworks aimed at analyzing and redesigning metabolic and signaling networks.

(1) Metabolic Network Gap Filling: Existing stoichiometric metabolic reconstructions, even for well studied organisms such as *E. coli*, include "unreachable" or blocked reactions due to the inherently incomplete nature of the reconstructed metabolic maps. These blocked reactions cannot carry flux under any uptake conditions. In this project we first identify all such blocked reactions and subsequently pinpoint which reactions to add to the existing model to bridge the maximum number of such gaps. The minimal set that accomplishes this task is chosen from an encompassing list of candidate reactions constructed from databases such as KEGG and Metacyc. The developed framework is demonstrated on genome-scale metabolic models of *Escherichia coli* and *Saccharomyces cerevisiae.* Reactions with higher BLAST scores against the genome of the curated model are preferentially selected. In addition, information as to which metabolites are present (e.g., CE-MS measurements) can be integrated into the gap-filling procedure.

(2) Assessing Objective Functions Driving Metabolic Responses to Perturbations: Genome-scale metabolic reconstructions are increasingly being used to predict the response of metabolic networks to genetic (e.g., gene knock-outs) and/or environmental (e.g., high/low glucose) perturbations. This is accomplished by optimizing an objective function that abstracts the dominant factors driving flux reallocation. These postulated hypotheses include biomass formation maximization, minimization of metabolic adjustment (MOMA)[1] , regulatory on/off minimization (ROOM) [2], etc. In this project, we assess the quantitative performance of these hypothesized objective functions in response to genetic and/or environmental perturbations and propose a new one based on flux ratios rather than

* Presenting author

absolute values. A comprehensive comparison using experimental data for wild-type and perturbed networks alludes to the use of composite objective functions as the best predictors.

(3) Elucidating Fluxes in Genome-scale Models Using Isotopomer Labeling Experiments: Isotopic label tracing is a powerful experimental technique that can be combined with the constraint-based modeling framework to quantify metabolic fluxes in underdetermined systems. The calculation of intracellular fluxes by 13C-MFA is based on the fact that when cells are fed a growth substrate with certain carbon positions labeled with 13C, the distribution of this label in the intracellular metabolites can be precisely determined based on the known biochemistry of the participating pathways. Most labeling studies focus on skeletal representations of central metabolism and ignore many flux routes that could contribute to the observed isotopic labeling patterns. In addition, often times a wide range of flux values could explain the experimentally observed labeling patterns in network areas where the experimental measurements provide low resolution. In this work, we investigate the importance of carrying out isotopic labeling studies at the genome-scale. Specifically, we explore how the activity of multiple alternative pathways could in many cases adequately explain the experimentally measured labeling patterns and also suggest methods for improving the resolution of quantified fluxes. Finally, we investigate the effects of introducing global metabolite balances on cofactors such as ATP, NADH, and NADPH as their inclusion in labeling analysis is often neglected but may be important for obtaining biologically realistic flux distributions.

(4) Optimal Redesign: Our research group developed the OptKnock[3] and Optstrain[4] procedures for microbial strain redesign through targeted gene additions and deletions. Both procedures use the maximization of biomass to predict flux reallocations in the face of genetic perturbations. Here we will present how to extend these optimization frameworks to account for popular quadratic objective functions such as MOMA[1] and contrast the obtained results. In addition, we will discuss how to computationally integrate modulations (i.e., up or down gene regulations) in addition to knock-in/outs in the palette of allowed genetic manipulations for microbial strain optimization[5].

(5) Signaling Networks: The same pathway modeling concepts that have been extensively applied to analyze and optimize metabolite flows in metabolic networks can also be used to analyze and redirect information flow in signaling networks. Here we describe optimization-based frameworks for elucidating the input-output structure of signaling networks and for pinpointing targeted disruptions leading to the silencing of undesirable outputs while preserving desirable ones. The frameworks are demonstrated on a large-scale reconstruction of a signaling network composed of nine signaling pathways. Results reveal that there exist two distinct types of outputs in the signaling network that either can be elicited by many different input combinations or are highly specific requiring dedicated inputs. Furthermore, identified targeted disruptions are not always in terminal steps. Many times they are in upstream pathways that indirectly negate the targeted output by propagating their action through the signaling cascade.

References

1. Segre D, Vitkup D, Church GM (2002) “Analysis of optimality in natural and perturbed metabolic networks,” *PNAS* 99: 15112-15117.
2. Sholmi T, Berkman O, Ruppin E (2005) “Regulatory on/off minimization of metabolic flux changes after genetic pertubations,” *PNAS* 102: 7695-7700.
3. Burgard AP, Pharkya P, Maranas CD (2003) “OptKnock: A Bilevel Programming Framework for Identifying Gene Knockout Strategies for Micorbial Strain Optimization,” *Biotechnology and Bioengineering* 84: 647-657.
4. Pharkya P, burgard AP, Maranas CD (2004) “OptStrain: A Computational Framework for Redesign of Microbial Production Systems,” *Genome Research* 14: 2367-2376.
5. Pharkya P, Maranas CD “An optimization framework for identifying reaction actvation/inhibition or elimination candidates for overproduction in microbial systems,” *Metabolic Engineering* (*In press*).

* Presenting author

93

Next Generation Computational Tools for Biochemical Network Models

Ravishankar Rao Vallabhajosyula, Frank Bergmann, and **Herbert M. Sauro*** (hsauro@kgi.edu)

Keck Graduate Institute, Claremont, CA

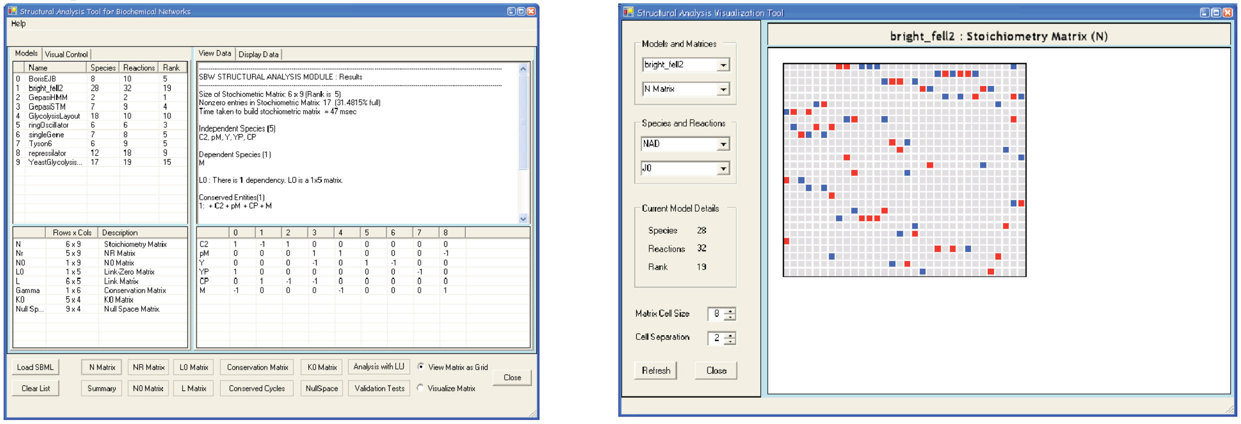
Introduction

In this contribution we wish to introduce three developments that will assist in the simulation of large scale biochemical networks that are of interest to the GTL community. Firstly we have developed a new algorithm to compute the conserved moieties of large biochemical networks. Secondly we have developed a new PC based high speed simulator that is at least an order of magnitude faster than current simulators. Finally we wish to describe a new real-time visualization tool that permits users to view simulations of biochemical systems in a more natural and meaningful way.

Conservation Analysis of Large Biochemical Networks

Conservation Analysis of biochemical models is an important step in determining the dependent and independent species in a biochemical network leading to a dimensional reduction of the model. This is a numerically intensive task that becomes error-prone with existing methods for large networks. The computation of the correct conserved cycles is also essential for the computation of the reduced Jacobian, which is non-singular. This reduced Jacobian is very important for a number of subsequent analyzes such as Bifurcation Analysis[1] and Metabolic Control Analysis[2].

Our new method makes uses of the Householder QR factorization of the stoichiometric matrix to obtain the relevant conservation relations for biochemical networks. Its advantage lies in much greater enhanced reliability and accuracy over other methods currently in use. The underlying algorithm is described in detail in our recent paper[4]. It has been integrated into Systems Biology Workbench (SBW)[3] and accepts models in standard SBML. SBW integration allows other developers to easily access the capability of this method. A separate graphical interface for this tool is also provided (See Figure 1).



	Name	Species	Reactions	Rank
0	BorisEJB	8	10	5
1	bright_fell2	28	32	19
2	GepasiHMM	2	2	1
3	GepasiSTM	7	9	4
4	GlycolysisLayout	18	10	10
5	ringOscillator	6	6	3
6	singleGene	7	8	5
7	Tyson6	6	9	5
8	repressilator	12	18	9
9	YeastGlycolysis...	17	19	15

	Rows x Cols	Description
N	6 x 9	Stoichiometry Matrix
Nr	5 x 9	NR Matrix
N0	1 x 9	N0 Matrix
L0	1 x 5	Link-Zero Matrix
L	6 x 5	Link Matrix
Gamma	1 x 6	Conservation Matrix
K0	5 x 4	K0 Matrix
Null Sp...	9 x 4	Null Space Matrix.

```
--------------------------------------------------------------------------------
SBW STRUCTURAL ANALYSIS MODULE : Results
--------------------------------------------------------------------------------
Size of Stochiometric Matrix: 6 x 9 (Rank is  5)
Nonzero entries in Stochiometric Matrix: 17  (31.4815% full)
Time taken to build stochiometric matrix  = 47 msec

Independent Species (5)
C2, pM, Y, YP, CP

Dependent Species (1)
M

L0 : There is 1 dependency. L0 is a 1x5 matrix.

Conserved Entities(1)
1:  + C2 + pM  + CP + M
```

	0	1	2	3	4	5	6	7	8
C2	1	-1	1	0	0	0	0	0	0
pM	0	0	0	1	1	0	0	0	-1
Y	0	0	0	-1	0	1	-1	0	0
YP	1	0	0	0	0	0	0	-1	0
CP	0	1	-1	-1	0	0	0	0	0
M	-1	0	0	0	-1	0	0	0	1

Models and Matrices: bright_fell2; N Matrix

Species and Reactions: NAD; J0

Current Model Details

Species	28
Reactions	32
Rank	19

Matrix Cell Size: 8

Cell Separation: 2

Figure 1: A Graphical tool for conservation analysis of biochemical network models

* Presenting author

High-Performance Simulation: roadRunner simulator

One of our aims is to create a rich interactive user experience for the modeling of biochemical systems. A critical requirement for this is the availability of high performance simulation software. The bulk of existing simulation tools rely on an interpretative method for evaluating the model equations which in turn is slow and inefficient. Instead we have been experimenting with just-in-time compilation of models using Java and .NET. A new high-performance simulator, codenamed 'roadRunner', drives this effort. This simulator uses the conservation analysis via SBW to construct the reduced model. It also implements most of the features specified in SBML level 2, including discrete events and user defined functions. This simulator has been developed using C# and relies on Sundials CVODE and NLEQ. The powerful reflection features of C# together with on-the-fly code generation allow 'roadRunner' to outperform a Java implementation. Initial tests also show an order of magnitude increase in performance over existing tools such as Jarnac and SBML odesolver. In addition to simulation, roadrunner also implements many other analyzes such as metabolic control analysis, model fitting and frequency analysis.

3D Time-Course Visualization of Simulations

Modeling and analysis of reaction-networks relies heavily on simulation tools. Traditionally the simulation results are available either as data tables or X-Y plots. Data tables are helpful for further processing by other computational tools. X-Y plots however tend to get complex even for a limited number of species. In creating a new visualization tool we had two goals in mind. The first goal was to tie the simulation results closely to the model and the other was to enable the user to view the simulation in real time.

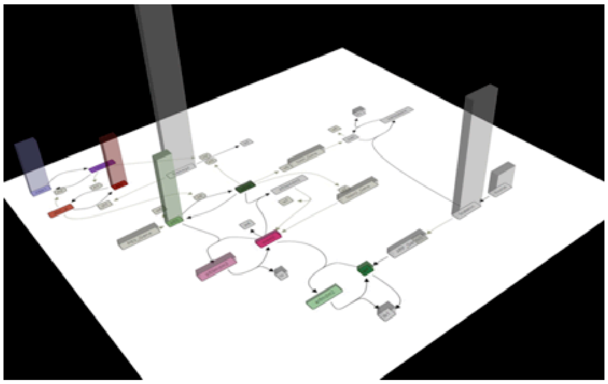


Figure 2: The 3D visualization tool showing simulation data in real-time

SBML is an established standard for the exchange of model-information. Recently the SBML group adopted a layout standard. This means it is now possible to store information about species, compartments and positions or dimensions of the model. This has allowed us to make the layout of a model portable. This layout implements the basis of a 3D visualization tool and is projected onto a 3D plane. Furthermore all positions of species are recognized and their concentrations are rendered as columns on top of the 3D plane. The heights of these columns represent the current species concentration values that vary during the time-course simulation. Since the time course simulation is performed continuously it is possible to dynamically focus on specific aspects of the model (see Figure 2).

The simulator and the visualization tool that implements the conservation analysis described earlier will be showcased at the DOE GTL meeting, and the results presented in a poster. A test version of the 3D Time-course Visualization can be found under: http://public.kgi.edu/~fbergman/Simulate3D.htm. The software for the latest version of Systems Biology Workbench is freely available at http://www.sys-bio.org.

References

1. V. Chickarmane, S.R. Paladugu, F. Bergmann and H.M. Sauro. "Bifurcation discovery tool," *Bioinformatics*, 21(18), 3688-3690, (2005).

* Presenting author

2. J-H.S. Hofmeyr. Metabolic control analysis in a nutshell, *Proc. Intl. Conf. Systems Biology*, Pasadena, California, 291-300, (2000).

3. H.M. Sauro, M. Hucka, A. Finney, C. Wellock, H. Bolouri, J. Doyle, and H. Kitano. Next generation simulation tools: the Systems Biology Workbench and BioSPICE integration. *OMICS* Winter, 7(4), 355-372, (2003).

4. R.R. Vallabhajosyula, V. Chickarmane, and H.M. Sauro. Conservation analysis of large biochemical networks. *Bioinformatics Advance Access* published on November 29, 2005, DOI 10.1093/bioinformatics/bti800.

94

Cyclic AMP-Dependent Regulatory Networks of *Shewanella oneidensis* MR-1 Involved in Anaerobic Energy Metabolism

Alex S. Beliaev[1]* (alex.beliaev@pnl.gov), Daad A. Saffarini[2], Lee Ann McCue[1], Amoolya H. Singh[3], Yang Zhang[1], Matthew J. Marshall[1], Grigoriy E. Pinchuk[1], and Jim K. Fredrickson[1]

[1]Pacific Northwest National Laboratory, Richland, WA; [2]University of Wisconsin, Milwaukee, WI; and [3]University of California, Berkeley, CA

Shewanella oneidensis MR-1 is a facultative metal-reducing bacterium with extensive respiratory versatility. Unlike many bacteria studied to date, the ability of *S. oneidensis* to grow anaerobically with several electron acceptors is regulated by the cAMP-receptor protein (CRP). CRP-deficient mutants of MR-1 are impaired in anaerobic reduction and growth with Fe(III), Mn(IV), fumarate, nitrate, and DMSO. Loss of anaerobic respiration in Crp$^-$ mutants is due to loss of terminal anaerobic reductases and not due to deficiency in carbon metabolism. To further elucidate the role of CRP and to understand the mechanisms of cAMP-dependent gene expression under anaerobic conditions in *S. oneidensis* MR-1, a combination of experimental and computational approaches have been applied.

To study the evolution of Crp in *S. oneidensis* MR-1 and its functional divergence from other closely related γ-*Proteobacteria*, we examined the conservation of the *crp* coding region as well as its upstream promoter region. Early results show that the coding region is conserved to almost 100% identity in *S. oneidensis* MR-1 and five other species of *Shewanella* (*S. amazonensis* SB2B, *S. baltica* OS155, *S. denitrificans* OS-217, S. *frigidimarina* NCIMB 400, and *Shewanella* sp. PV-4), and conserved to 95-97% identity with several closely related γ -*Proteobacteria* (*Escherichia*, *Salmonella*, *Shigella*, *Yersinia*, and *Vibrio*). Outside of this clade, *crp* appears to have diverged significantly in sequence since the last common ancestor. The ratio of synonymous to non-synonymous nucleotide substitutions (dN/dS, or K_a/K_s) of *crp* in these same species indicates that the gene is under strong selective pressure not to undergo mutation.

Microarray analyses mRNA expression profiles of wild-type and crp mutant cells grown anaerobically with different electron acceptors indicated that CRP positively regulates the expression of genes involved in energy generation and transcriptional regulation. These include the periplasmic nitrate reductase (*napBHG*), the polysulfide reductase (*psrAB*), anaerobic DMSO reductase (*dmsAB*) genes, as well as the nitrate/nitrite sensor protein *narQ*. Mobility shift assays using purified CRP suggest that this protein activates gene expression directly by binding to promoter regions of anaerobic reductase genes. Furthermore, our experiments indicate that cAMP is required for CRP activation. Mobility shifts were observed only when cAMP was added to CRP-DNA reaction mix and cAMP

* Presenting author

addition to aerobically growing *S. oneidensis* cells resulted in the induction of fumarate and Fe(III) reductase activities.

The genome sequence of *S. oneidensis* MR-1 contains three putative adenylate cyclase genes, designated *cyaA*, *cyaB*, and *cyaC*. Deletions of both *cyaA* and *cyaC* resulted in anaerobic growth deficiency with DMSO, nitrate, Fe(III), Mn(IV), and fumarate. These phenotypes are similar to the phenotypes of the CRP-deficient mutants. The function of both *CyaA* and *CyaC* as adenylate cyclases was confirmed by complementing an *E. coli* cyaA mutant. It is interesting to note that *CyaC* contains a predicted membrane-bound domain, similar to eukaryotic adenylate cyclases that are involved in signaling. One hypothesis to be tested is that the membrane domain of *CyaC* is involved in oxygen sensing, and therefore cAMP synthesis by this protein occurs under anaerobic conditions. *CyaC* consists of a membrane domain and a catalytic domain that is predicted to reside in the cytoplasm. Surprisingly, deletion of the *CyaC* membrane domain leads to loss of enzyme activity, suggesting that it may play a role in the stability or activation of the catalytic domain. Further work to identify the cAMP/CRP -dependent regulatory networks in *S. oneidensis* MR-1 is underway.

95 MEWG

Engineering *E. coli* to Maximize the Flux of Reducing Equivalents Available for NAD(P)H-Dependent Transformations

Patrick C. Cirino* (cirino@engr.psu.edu), Costas D. Maranas, and Jonathan W. Chin

Pennsylvania State University, University Park, PA

Biocatalysis offers the opportunity for unmatched reaction specificity and product diversity, and is integral to realizing a future of cost-effective "green" chemistry. The ever-growing ease with which we are able to manipulate cellular metabolism and to design enzymes with specified properties presents the opportunity to develop biocatalytic systems of increased complexity and efficiency. Transformations of primary importance are those catalyzed by reduced nicotinamide cofactor-dependent reductases, dehydrogenases and oxygenases. While significant improvements have been made in the biocatalytic properties of these enzymes, many issues remain unresolved regarding the preferred approach for implementing their transformations and the accompanying cofactor regeneration requirement. The primary objective of this research is to develop microbial production strains which will serve to host NAD(P)H-dependent, heterologous reactions with maximized efficiency in the generation and subsequent utilization of reduced cofactors derived from glucose or other renewable energy sources. *E. coli* is the microbial host chosen for these studies, and we are initially studying the reduction of xylose to xylitol (by heterologous xylose reductases with different cofactor preferences) to serve as an experimental platform that allows us to systematically characterize the individual and synergistic influences of select genetic modifications on strain performance, measured as the yield on xylitol produced (from xylose reduction) per glucose consumed as co-substrate. Maximizing this yield translates to uncoupling carbon metabolism from respiration or fermentation and effectively "respiring" on xylose.

In addition to studying metabolic parameters that are *expected* to impact our objectives (e.g., fermentative pathways, global regulators, transhydrogenase function), stoichiometric network analysis and the strain optimization framework OptKnock are being used to understand the influence of enzymes with potentially critical or ambiguous physiological functions on theoretical yields and to suggest knockout strategies that will constrain the network such that cell growth is coupled to

* Presenting author

xylitol production. For example, xylose transport via the ATP-dependent transporter was shown to significantly reduce xylitol yields compared to xylose transport via proton symport or facilitated diffusion. Furthermore, the physiological roles of the two native transhydrogenases (i.e., whether they are reversible) critically influence flux distributions and yields. Experimental results for wild-type and deletion strains are being used to better understand network architecture, supplement existing metabolic models, and improve prediction accuracy and fidelity.

Maximizing xylitol production inherently leads to very low growth rates because ATP yields must be low if reducing equivalents are to be directed towards xylose reduction. Consequently, deletion of ATP synthase is chosen by Optknock to be a key genetic modification required for coupling xylitol production to biomass formation. Deletion of *atpA* increased the xylitol yield by ~40% in shake-flask cultures (normalized to cell density). In order to more accurately reflect conditions of low growth, we are also using metabolically active but non-growing “resting cells” to evaluate strain performance. Use of resting cells additionally eliminates growth as a variable that alters partitioning of carbon and reducing equivalents, and provides a more reliable method for determining maximum experimental yield. The yield on xylitol per glucose consumed in our “base” strain expressing an NADPH-dependent xylose reductase improved from 1.5 in batch culture to 4.3 under non-growing conditions. Measuring these yield values from knockout strains enables us to illuminate the contributions of various pathways and reactions to NADPH-dependent xylose reduction, characterize the effects of overexpression of an NADPH sink (a common scenario in whole-cell biocatalysis) on partitioning of co-substrate (i.e., glucose) carbon and reducing equivalents, and identify combinations of genetic modifications that lead to improved strain efficiency and productivity.

96

A New MILP Based Approach for *in Silico* Reconstruction of Metabolic Networks and Its Application to Marine Cyanobacterium *Prochlorococcus*

Xiaoxia (Nina) Lin* (xiaoxia@genetics.med.harvard.edu), Aaron Brandes, Jeremy Zucker, and **George M. Church**

Harvard Medical School, Boston, MA

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

The increasing availability of annotated genomes has rendered the possibility of applying systems approaches, e.g. flux balance analysis (FBA), to the study of a large variety of organisms. To achieve this high-throughput goal, we have been developing an automatic bioinformatics pipeline to facilitate the process of generating *in silico* whole-cell metabolic models from genome annotations. In this work, we present a new computational framework for the key step in this pipeline which constructs metabolic networks by integrating genome annotation, reaction database, and phylogenetic information.

Our goal is to construct metabolic pathways/networks for a new species based on its genome annotation and a multiple-species pathway/reaction database (e.g. BioCyc databases). Using a mixed-integer linear programming (MILP) optimization framework, the new algorithm selects a set of reactions from a universal super-network which can achieve the functionality of a pathway or network to convert specific metabolites or to enable the cell to live and grow. The solution includes not only reactions already identified in the genome annotation but also additional ones required to achieve the

* Presenting author

functionality which are the most possible phylogenetically. Alternative and/or sub-optimal solutions can also be systematically generated to increase the likelihood of identifying the real biological network. In addition, quantitative data such as nutrient condition can be readily incorporated to improve the predictions.

The above approach has been applied to the study of a marine cyanobacterium *Prochlorococcus marinus*, which dominates the phytoplankton in the tropical and subtropical oceans and contributes to a significant fraction of the global photosynthesis. As a proof-of-concept, the algorithm automatically generated novel TCA pathways which do not exist in the pathway databases and are consistent with partial knowledge of cyanobacteria. We have also successfully reconstructed the networks for central carbon metabolism, amino acid biosynthesis, and nucleotide biosynthesis. We are currently moving towards the whole-genome metabolic network of *Prochlorococcus* as well as using the results from this approach to identify missing genes/enzymes and to refine the genome annotation.

97 MEWG

Optimizing Central Metabolic Pathways in Yeast

Thomas W. Jeffries[1,2]* (twjeffri@wisc.edu), Chenfeng Lu[1], Karen J. Mansoorabadi[1], Jennifer R. Headman[1], Ju-Yun Bae[1], and Haiying Ni[3]

[1]University of Wisconsin, Madison, WI; [2]USDA, Forest Products Laboratory, Madison, WI; and [3]University of Chicago, Chicago, IL

Optimization of metabolite flux relies primarily on deletion, overexpression, or regulated expression of one or more target genes with the assumption that these changes will be reflected in altered levels of enzymatic activities. Predictive models are essential, but in practice, genetic engineering must be carried out with empirical trials and concomitant modification of environmental variables. Targeted changes do not always have the desired effects.

Our research has concentrated on engineering xylose metabolism in yeasts. We explore the metabolism of *Pichia stipitis*, which is capable of converting xylose to ethanol, and we modify *Saccharomyces cerevisiae*, which is not. The two approaches are complementary. No one feature of *P. stipitis* has as yet proven to be the "magic key" that enables xylose fermentation. *S. cerevisiae* possesses orthologs of all the known steps for xylose assimilation (*GRE3, SOR1, XKS1*), but does not naturally coordinate their expression in a way that permits the conversion of xylose to ethanol. Overexpression of the corresponding *P. stipitis* genes (*XYL1, XYL2, XYL3*) can improve both assimilation and ethanol production in *S. cerevisiae*, but elevated expression of *XYL3* (or *XKS1*) can inhibit growth on xylose unless accompanied by the overexpression of either Ps or Sc*TAL1* (transaldolase).

We have conducted a systematic search for genes that will enhance growth on xylose when deleted or overexpressed. One unpredicted finding was *PHO13*, a *p*-nitrophenyl phosphatase that shows some specificity for histone dephosphorylation. Deletion of this gene relieves growth inhibition on xylose and increases *ScTAL1* expression approximately two-fold. Our overexpression studies have identified three additional genes that appear to relieve inhibition when expressed along with *XYL1, XYL2* and *XYL3*.

Coordinated expression – even at low levels – appears to be important, so we have developed a series of *S. cerevisiae* promoters that enable regulated expression over a dynamic range of about 80-fold. We have also developed methods that enable the relatively rapid switching of promoter/gene pairs to

* Presenting author

screen the effects of regulated expression with multiple genes simultaneously. The effects of multigene expression depend greatly on the environmental conditions and the physiological state of the cells such that one combination of promoters can be appropriate for one condition (e.g. high aeration), but less desirable for another (e.g. limited oxygen). Such effects are expected to extend to other regulatory signals as well.

As reported previously at this conference, heterologous expression in plants can be affected by gene orientation. We have explored this phenomenon and have concluded that gene orientation and the repeated use of a single promoter (promoter dilution) does not seem to have a significant effect in *S. cerevisiae.*

The genetic background can greatly affect results. As mentioned, mutations can be introduced – or they can arise spontaneously. To further explore this aspect, our research has examined the effects of *XYL1, XYL2,* and *XYL3* overexpression in several laboratory and industrial yeast strains.

98 MEWG

Multi-Scale Models for Gene Network Engineering

Yiannis N. Kaznessis* (yiannis@cems.umn.edu)

University of Minnesota, Minneapolis, MN

Armed with increasingly fast supercomputers and greater knowledge of the molecular mechanisms of gene expression, it is now practical to numerically simulate complex networks of regulated biological reactions, or gene circuits. It is also becoming feasible to calculate the free energy of noncovalent binding of regulatory proteins to specific DNA target sites. Using a hybrid stochastic-discrete and stochastic-continuous simulation algorithm, we obtain an accurate time-evolution of the behavior of complex gene circuits, including a clear picture on the role of highly dilute, but significant, regulatory proteins. These regulatory proteins are responsible for the non-linear control used by biological organisms to regulate their most important processes. The network simulations provide insight, which can guide rational engineering of regulatory proteins and DNA operator sequences using molecular mechanics simulations. In this presentation we examine two important gene circuits, the bistable switch and the oscillator. We study the role of specific biomolecular interaction phenomena on the dynamics of these gene circuits. Using models that span multiple time and space scales, from atomistic, to molecular, to interaction networks we develop design principles for high quality bistable switch and oscillator circuits.

99

Generalized Computer Models of Chemoheterotrophic Bacteria: A Foundation for Building Genome-Specific and Chemically-Detailed Bacterial Models

Jordan C. Atlas, Evgeni V. Nikolaev, and **Michael L. Shuler*** (mls50@cornell.edu)

Cornell University, Ithaca, NY

A significant challenge in systems biology is to better understand fundamental design principals of cellular organization and function by taking advantage of information encoded in annotated DNA sequences. While bioinformatics tools and related technology will continue to dominate within the next decade, these efforts are, by themselves, insufficient. New tools are necessary to explicitly relate genomic and molecular information to cellular physiology and population response. The release of the 1000th microbial genome is expected within next few years (Overbeek *et al.,* 2005), and this further accelerates the development of technology to build accurate metabolic reconstructions and balanced stoichiometric models of bacterial cells. Stoichiometry correctly defines overall barriers for intracellular steady-state fluxes under fixed 'defined medium' constraints, and genome-scale stoichiometric models have been very successful in many instances in fundamental and applied research (Reed and Palsson, 2003). However, the predictive capability of static stoichiometric models is limited to the calculation of 'instant snapshots' of different phenotypes and therefore such models cannot capture dynamic changes in protein machinery, metabolite concentrations, and cell geometry. The advent of abundant data coupled with the limitations of current modeling approaches necessitates the development of novel modeling frameworks to rapidly build completely functional bacterial cell models.

The availability of complete metabolic reconstructions for a variety of bacterial species can facilitate the development of molecularly-detailed bacterial subsystems models, named here 'modules.' Such *chemically detailed* modules can then be combined within a *coarse-grained* model to produce a completely functional *hybrid* single cell model (Castellanos *et al.,* 2004). The initial step of our 'hybrid model approach' was to construct a coarse-grained model with lumped 'pseudochemical species.' The coarse-grained model explicitly links DNA replication to metabolism, cell cycle, cell geometry and external environment (Browning, *et al.*, 2004). Such models can include known or putative regulation, relations capturing key events in the production and utilization of cell's energy and redox equivalents, RNA transcripts, proteins, lipids, and different forms of nucleotides. These coarse-grained models correctly predict sustained bacterial reproduction (*i.e.*, chromosome replication and cell division following one another in the right order and timing), which can be viewed as an *obligatory* test for all whole cell models. Modest-sized coarse-grained models allow for the application of rigorous mathematical analyses such as bifurcation and stability analyses to verify the model's robustness.

This hybrid-modular framework has been successfully applied to create Cornell's Minimal Cell Model (MCM). A 'minimal cell' is a hypothetical free living organism possessing the functions required for sustained reproduction in a maximally supportive culture environment. The 'modularity' has been demonstrated by constructing a genomically and chemically detailed model of nucleotide metabolism within the MCM (Castellanos *et al.,* 2004), utilizing statistical mechanics methods for parameter estimation (Brown and Sethna, 2003). We are currently in the process of incorporating detailed genome-specific information into the *E. coli* model. Our focus will be on the core carbon metabolic subsystems such as energy metabolism, nucleotide biosynthesis, and biomass precursor formation. The carbon metabolism provides 12 key precursors for all biochemicals formed within

* Presenting author

the bacterium and allocates larger fluxes of all intracellular fluxes unevenly distributed throughout metabolism. The 'core' hybrid model will allow us to gain fundamental insight into how the dynamic events in the central carbon metabolism are controlled by changes in the chromosome and external environment. The core model will also include an improved model of chromosome replication and lumped modules for subsystems with diminished fluxes. This modeling framework will serve as a platform for the development of a variety of bacterial cell models. Specifically, the approach will be applied to the development of a functionally complete model of *Shewanella oneidensis*, a microorganism closely related to *E. coli*. The constructed models will be publicly available as Matlab modules and via the System Biology Markup Language (SBML) format. The current project is conducted in the cooperation with Gene Network Science, Inc., and utilizes the GNS VisualCell™ modeling platform.

References

1. Brown, K.S., and J.P. Sethna. 2003. "Statistical Mechanics Approaches to Models with Many Poorly Known Parameters." *Phys. Rev. E* 68, 021904.
2. Browning, S.T., M. Castellanos, and M.L. Shuler. 2004. "Robust Control of Initiation of Prokaryotic Chromosome Replication: Essential Considerations for a Minimal Cell." *Biotechnol. Bioeng.* 88(5):575-584.
3. Castellanos, M., D.B. Wilson, and M.L. Shuler. 2004. "A Modular Minimal Cell Model: Purine and Pyramidine Transport and Metabolism." Proc. Natl. Acad. Sci. (USA). *PNAS* 101(17): 6681-6686.
4. Overbeek, R., T.Begley *et al* 2005. "The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes," *Nucleic Acids Res.* 33(17): 5691-5702
5. Reed, J.L. and Palsson, B.Ø. 2003. "Thirteen Years of Building Constraint-Based *in silico* Models of *Escherichia coli*," *Journal of Bacteriology*, 185(9): 2692-2699.

100

Determination of the Most Probable Objective Function for Flux Analysis of Metabolism Using Bayesian-Based Model Discrimination

Andrea Knorr and **Ranjan Srivastava*** (srivasta@engr.uconn.edu)

University of Connecticut, Storrs, CT

Metabolic flux analysis has proven to be a powerful tool for analyzing, understanding, and engineering a variety of different organisms. However, when carrying out flux analysis, the metabolic reaction network is often underdetermined due to a lack of information. Under such circumstances, optimization theory may be used to determine the fluxes across the metabolic pathways. To carry out the optimization process, an appropriate objective function must be identified. Ideally the objective function should represent a biological process which, when optimized, would prove more beneficial for the organism than the optimization of any other process. Several such objective functions have been proposed, including maximization of growth and optimization of energy efficiency. In this work, we have adapted a Bayesian-based model discrimination method to allow us to determine which objective function is most probable for use with flux analysis.

The model system analyzed was the central metabolism of *E. coli* growing on succinate. Based on biological plausibility, two objective functions were compared. The first was maximization of growth, while the second was minimization of redox potential. Flux balance analysis and linear programming coupled with experimental data were used to predict metabolic fluxes using each objec-

* Presenting author

tive function. Our adapted Bayesian model discrimination approach was then employed to determine which of the two objective functions was more likely. Growth maximization was shown to be the most probable objective function for use with flux analysis to determine metabolite distribution. It should be noted that the technique employed here is generic enough to be used to determine the most probable objective function for the metabolic analysis of any organism.

101

In silico Analysis of *Escherichia coli* Metabolism to Optimize Electron Production

David Byrne* (dbyrne@bu.edu), Daniel Segre, and **Timothy Gardner**

Boston University, Boston, MA

In metabolic engineering, it is desirable to manipulate microbes to overproduce specific by-products. In order to attain this bioengineering objective it would be ideal to be able to use informed computational predictions to effectively direct subsequent experimental microbial biotransformations. We are specifically interested in developing methods to predict environmental conditions or mutant microbial strains that optimize the electron production for potential application in microbial fuel cells and bioremediation.

To achieve this objective, the methods that we are developing incorporate simulations of genome-scale metabolic models, initially that of the *Escherichia coli* K-12 MG1655 consisting of 931 unique reactions and 625 metabolites. An algorithm searches these simulations for optimal environmental conditions and genetic modulations that maximize electron yield while minimizing competing cellular processes.

This problem may be mathematically posed as a bi-level optimization problem such that the inner optimization problem performs a linear programming optimization for maximum biomass yield based on stoichiometric and nutrient constraints while the outer problem maximizes electron production by modulating specific reactions available to the inner problem. A computational optimum may then be attained by simultaneously solving both the inner and outer problems using multiple-integer linear programming.

This computational framework may be used to address the following questions of biological significance: How does electron production relate to energy (e.g. ATP synthesis) and growth yield? How does electron production relate to changes in a microbe's environment? How does it relate to the general conditions requiring aerobic respiration, anaerobic respiration, or fermentation? How does electron production relate to specific changes in carbon, nitrogen, electron-acceptor, or other media substrate sources? How does electron production change when different types of genetic perturbations are imposed?

* Presenting author

102

Reverse-Engineering the Central-Metabolism Network of *Shewanella oneidensis* MR-1

J.F. Penders and **T.S. Gardner*** (tgardner@bu.edu)

Boston University, Boston, MA

With the recent availability of genome annotations for many microbial organisms, metabolic engineering has seen the development of genome-scale metabolic models. This new engineering tool seems promising for optimizing microbes for bioremediation and microbial fuel cell applications. Towards this metabolic engineering objective, the process of developing the most accurate possible model is essential. Current methods are based on intuition and consist in adding or removing reactions to improve model agreement with experimental data.

Shewanella exhibits an uncommon and complex metabolism. A facultative anaerobe, it has versatile respiratory pathways, and is able to use a broad spectrum of electron acceptors, including metals such as iron, manganese and uranium. In preliminary work, we have applied current methods to develop a metabolic model for the central metabolism of *Shewanella oneidensis* MR-1. We created a Pathway-Genome Database, containing 206 pathways, 1015 reactions and 807 enzymes using existing genome annotations coupled with manual curation. We used this metabolic map to build a model for the central-metabolism of *Shewanella*, but our model predictions diverged from known experimental observations kindly shared with us by the *Shewanella* Federation.

Shewanella's complex respiratory metabolism thus complicates intuitive approaches about which reactions should be added to or removed from the model. Therefore, although current methods based on manual tuning are efficient for most organisms, they risk becoming cumbersome and are inappropriate for organisms exhibiting non-intuitive metabolism. In such cases, an automatic way to learn the metabolic network from experimental data is needed.

We have thus developed a novel computational method tool to reverse-engineer metabolic networks in *Shewanella*. Metabolic networks have unique properties such as high connectivity of certain nodes (such as co-factors) and non-pairwise connectivity of edges, and they require multiple-edge perturbations for effective network inference approaches. Our algorithm semi-greedily searches all biochemically feasible metabolic networks, and learns an optimal network by iterative comparison of experimental data with computed metabolite producibility and flux balance analysis.

We hope that the application of our novel method to *Shewanella* will aid investigators in unraveling its central metabolism and provide new insights into its extraordinary reducing capabilities.

* Presenting author

103

Genome-Scale Metabolic Model of *Shewanella oneidensis* MR1

Brian Gates[1]* (bgates@genomatica.com), Grigoriy E. Pinchuk[2], Christophe Schilling[1], and Jim K. Fredrickson[2]

[1]Genomatica, Inc., San Diego, CA and [2]Pacific Northwest National Laboratory, Richland, WA

A genome-scale metabolic model of *Shewanella oneidensis* MR1 was created using the constraints-based approach. The development of the model was accelerated by an early stage prototype of an automated network reconstruction process that leveraged information contained in high-quality models of other organisms to predict a large percentage of the *Shewanella* metabolic network based on gene sequence homology. Subsequent manual network reconstruction efforts proceeded through review of the genome annotation and testing for the presence of certain essential pathways. The model was then refined in an iterative procedure in which model predictions were compared to experimental data.

The current model version includes 742 of the 5102 genes in *S. oneidensis*, which are translated to 626 proteins that catalyze 702 model reactions. An additional 33 reactions are present in the model without genetic support, referred to as non-gene associated reactions. An aggregate reaction for biomass formation was developed that included as substrates the major molecular components of biomass in the stoichiometry necessary to produce 1 g dry cell weight as well as sufficient ATP to meet the energetic demands of growth. Because detailed composition data for MR1 is not yet available, all simulations were performed using the biomass composition of *E. coli*. The model is currently able to simulate several aspects of metabolism in *S. oneidensis*, including substrate utilization profiles, byproduct secretion under certain conditions, and the use of several terminal electron acceptors. Consistent with experimental observations, the model predicts that growth on lactate under oxygen limited conditions will produce near-equimolar acetate secretion and no formate production. The model also predicts the known ability of MR1 to grow using iron, fumarate, or nitrate as a terminal electron acceptor.

As part of the on going efforts to continue refining the model to make it more consistent with the known physiology of the organism, we analyzed phenotypic respiration data generated using the Biolog Phenotype MicroArray™ platform for *S. oneidensis*. The array data was used to test the model's ability to predict aerobic respiration on 139 different carbon sources. Simulation matched experimental predictions for 29% (7/24) of the compounds that produced respiration, and 84% (97/115) of the compounds that did not. Further review of the genome led to the assignment of 6 additional metabolic reactions to genes, improving the respiration prediction rate to 42% (10/24).

The metabolic network of the *Shewanella* model was compared to the network of the *E. coli* genome scale model, the closest homolog used in the initial automated reconstruction. Genome-scale deletion analyses were performed with both models to determine essential genes for aerobic growth on lactate. 171 genes linked to 219 reactions were found to be essential in MR1, while 176 genes linked to 216 reactions were found in *E. coli* under the same conditions. However, there were 14 reactions that were essential only in MR1, and 10 in *E. coli*. Amino acid metabolism had more unique genes in *E. coli* than in MR1, due to the incorporation of serine glyoxylate aminotransferase as an alternate pathway for serine utilization in MR1.

All of the model development and simulation was enabled by the SimPheny™ software platform. The model and the SimPheny server dedicated to *S. oneidensis* is now being accessed by multiple

* Presenting author

distributed research groups in the *Shewanella* Federation as part of a beta-test for coordinated high quality remote access to SimPheny. Through the availability of a first version model and remote access to SimPheny we are beginning to enable model-driven discovery research for *S. oneidensis* in a collaborative, multi-institutional research setting. We expect that these efforts will accelerate the pace of discovery and our overall understanding of metabolic physiology in *Shewanella* species.

104

A Phylogenetic Gibbs Sampler for High-Resolution Comparative Genomics Studies of Transcription Regulation

Sean Conlan[1]* (sconlan@wadsworth.org), William Thompson[4], Lee Newberg[1,2], Lee Ann McCue[3], and **Charles Lawrence**[4]

[1]The Wadsworth Center, Albany, NY; [2]Rensselaer Polytechnic Institute, Troy, NY; [3]Pacific Northwest National Laboratory, Richland, WA; and [4]Brown University, Providence, RI

A thorough knowledge of an organism's transcription regulatory network is a critical component of understanding the biology of the organism as a whole. The foundation of a prokaryotic regulatory network is the *cis* (transcription factor binding sites) and trans (transcription factors) elements which constitute the molecular wiring diagram. Comparative genomics, which makes inferences about the properties of an organism through analysis of related genomes, offers a powerful set of tools for deciphering this network. The field of comparative genomics, particularly with regards to microbial genomics, is entering an unprecedented time, with high-throughput sequencing facilities enabling the sequencing of many closely related bacterial strains and isolates. By comparing closely-related genomes, there is the potential to discover the minor sequence changes responsible for important phenotypic variations observed between related bacterial strains. Using closely related strains for comparative studies of transcription regulation is attractive because these species are most likely to share common transcription factors, and therefore common *cis*-regulatory elements. Unfortunately, the recent speciation of closely related genomes results in correlation among the sequences that complicates the detection of functionally conserved motifs. To facilitate high-resolution comparative genomics studies, that are able to leverage the power of both closely- and distantly- related genomes, we have developed a phylogenetically-rigorous Gibbs recursive sampler, orthoGibbs. Phylogeny is incorporated through the use of an evolutionary model and sequence weights. OrthoGibbs was used to detect known transcription factor binding sites upstream of a study set of genes from *Escherichia coli* and 7 other gamma-proteobacterial genomes. We were able to demonstrate improved specificity and positive predictive value for orthoGibbs when compared to the non-phylogenetic Gibbs sampler. In addition, orthoGibbs identified sites known to be bound by Fur, upstream of iron-regulated genes in 11 sequenced *Shewanella* species.

* Presenting author

105

Integrated Computational and Experimental Approaches to Facilitated Model Development

R. Mahadevan* (rmahadevan@genomatica.com), T. Fahland, Y. Kwan, J.D. Trawick, I. Famili, and C.H. Schilling

Genomatica, Inc., San Diego, CA

Advances in high-throughput technologies have lead to the sequencing of several microbial genomes with important applications in bioremediation, CO_2 sequestration, alternative energy sources and industrial biotechnology. The availability of such sequence information has enabled the development of metabolic models, which are genomically and biochemically structured databases and are valuable for data interpretation and computational analysis. The models will also enable the rational design of strategies for improving the efficiency of bioremediation and CO_2 sequestration and other processes related to the Department of Energy's core missions. Genome-scale constraint-based modeling has been shown to be successful in predicting physiology under varied conditions. The manual development of such genome-scale models is labor intensive and time consuming. Thus, there is a critical need to develop approaches for the rapid development of systems level cellular models. In this project, we have developed an automated approach to develop genome scale models based on a combination of genome sequence and high-throughput phenotyping, that can be used to deliver a model-driven approach to biological discovery in the newly sequenced microorganisms. The initial phase of this project focuses on the design and prototype development of approaches for automated metabolic reconstruction based on both the genome sequence and through comparison with the existing high quality metabolic models in our database. We are also developing methods for automatically identifying candidate reaction sets to close gaps in the metabolic pathways associated with the synthesis of essential biomass components utilizing high-throughput experimental data from growth phenotyping experiments. Initial results from the proposed approaches will be shown for a newly developed genome-scale model of *Bacillus subtilis*.

106

New Technologies for Metabolomics

Jay D. Keasling* (jdkeasling@lbl.gov), Carolyn Bertozzi, Julie Leary, Michael Marletta, and David Wemmer

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

* Presenting author

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.

107

Metabolomic Functional Analysis of Bacterial Genomes

Clifford J. Unkefer[1]* (cju@lanl.gov), Pat J. Unkefer[1], Munehiro Teshima[1], Kwasi Godwin Mawuenyega[1], Norma H. Pawley[1], Rodolfo A. Martinez[1], Marc A. Alvarez[1], Daniel J. Arp[2], Luis Sayavedra-Soto[2], Xueming Wei[2], and Norman Hommes[2]

[1]Los Alamos National Laboratory, Los Alamos, NM and [2]Oregon State University, Corvallis, OR

Achieving the GTL goal of obtaining a complete understanding of cellular function requires integrated experimental and computational analysis of genome, transcriptome, proteome as well as the metabolome. Metabolite concentration is a product of cellular regulatory processes, and thus the metabolome provides a clear window into the functioning of the genome and proteome. 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. The power of metabolome analysis is greatly enhanced by stable isotope labeling. By combining stable isotope labeling and NMR/Mass spectral analysis, one can assess not only the metabolite concentration, but the flux through metabolic pathways as well. This approach is essential to establishing precursor product relationships, and to test if putative pathways identified from analysis of the genome are operational. Our initial results on the metabolome of *Nitrosomonas europaea* are reported here. *N. europaea* is a chemolithotroph, which derives its energy for growth from the oxidation of NH_3 to NO_2^- and fixes CO_2 as a carbon source.

Before sequencing its genome, *N. europaea* is only known to grow when provided with ammonia as an energy source, O_2 as the terminal electron acceptor, and CO_2 as the carbon source. However, *N. europaea* is not abstinent with regard to other energy sources, electron acceptors, and carbon sources. Based on homology analysis of the *N. europaea* genome genes encoding for a PTS fructose/mannose transporter, a complete glycolytic pathway, a complete TCA cycle, and a complete electron transport pathway. Based on the analysis of the genome, we have demonstrated that while using ammonia as an energy source, *N. europaea* can grow using a number of heterotrophic carbon sources. We cultured the organism in medium containing [1,2-$^{13}C_2$]pyruvate, and analyzed the ^{13}C-isotopomers of metabolites by MS and NMR spectroscopy. The results demonstrate that despite the fact that their genome encodes for a complete TCA cycle, *N. europaea* expresses only a branched TCA cycle in the presence of pyruvate.

Chemostat development - Essential to the comparative metabolomics, metabolic flux analysis and metabolic regulation studies are steady state culture conditions. During the ammonia oxidation process, the *N. europaea* acidifies its medium and builds up nitrate rapidly during growth of batch cultures. These conditions, coupled with the increasing cell density in exponential batch cultures make impossible to attribute the effect of a particular variable on the metabolite profiles or metabolic flux. To minimize culture variability, we have developed a chemostat specifically for growth of *Nitrosamonas europaea*. Our chemostat tightly regulates rate of addition of O_2, NH_3 and CO_2 using computer controlled mass flow controllers. In addition, the chemostat monitors and adjusts pH, [O_2], stirring rate, and the rate of nutrient addition. Our initial chemostat cultures of *N. europaea* were carried out at a 0.026 h^{-1} dilution. This culture was feed CO_2 and a nutrient solution at a constant rate. In this culture, the addition of ammonia was used to maintain the pH of the cultures. After two days, the culture reached steady state indicated by a constant $OD_{600nm} = 0.45$. In addition, the accumulated concentration of nitrate also plateaued indicating that the rate of production of nitrate by the culture was constant. As *Nitrosomonas europaea* grows it oxidizes ammonia. The linear rate of ammonia addi-

* Presenting author

tion is also an indication of steady state growth. We routinely achieve steady state cultures with an optical density (600nm) of 0.55, which is 3-5 times greater then that reported for any batch culture conditions. Initial metabolite profiles of these steady state cultures will be reported.

108

Linking Nano-Scale Patterns of ^{15}N and ^{13}C Metabolism to Bacterial Morphology

Jennifer Pett-Ridge[1]* (pettridge2@llnl.gov), Stewart J. Fallon[1], Juliette Finzi[2], Doug Capone[2], Radu Popa[2], Ken Nealson[1], Ian D. Hutcheon[1], and **Peter K. Weber**[1]

[1]Lawrence Livermore National Laboratory, Livermore, CA and [2]University of Southern California, Los Angeles, CA

The technical challenges of tracing isotopes within individual bacteria and nanoparticles are overcome with high resolution Nano-Secondary Ion Mass Spectrometry (NanoSIMS). Samples are sputtered with an energetic primary beam, liberating secondary ions that are separated by a mass spectrometer and detected in a suite of electron multipliers. Using this technique, five isotopic species may be analyzed concurrently with spatial resolution as fine as 50nm and isotope ratio precision of ± 1.5‰. A high sensitivity isotope ratio 'map' can then be generated for the analyzed area.

We used this technique to quantitatively describe ^{13}C and ^{15}N uptake and transport in two marine cyanobacteria grown on $NaH^{13}CO_3$ and $^{15}N_2$. These diazotrophic bacteria are faced with the

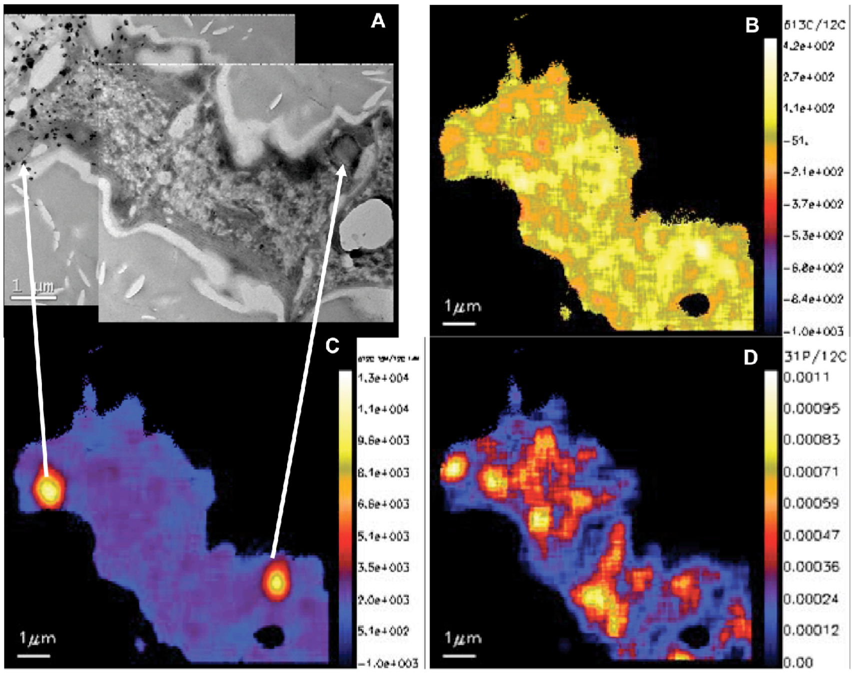


Figure 1. Corresponding NanoSIMS and TEM images of two Trichodesmium cells after 8 hrs of incubation with $^{13}C\text{-}HCO_3^-$ and $^{15}N\text{-}N_2$: (a) TEM, (b) C isotope ratio and (c) N isotope ratio (d) P: ^{12}C ratio. The filament was embedded in epoxy, ultramicrotomed into 200 nm thick sections, stained to reveal ultrastructure in TEM, and then analyzed in the NanoSIMS. The nitrogen image allows the TEM and NanoSIMS images to be correlated. The carbon and nitrogen isotope ratio data are shown as deviations from standard values in parts per thousand, as indicated in the legends ($\delta^{13}C$ and $\delta^{15}N$). ^{15}N enrichment (c) shows the localization of the newly fixed nitrogen. This fixation is correlated with N storage structures (cyanophycin) indicated in the TEM image.

* Presenting author

challenge of isolating regions of N-fixation (O_2 inhibited) and photosynthesis (O_2 producing). In *Anabaena oscillarioides*, we found that specialized N-fixing heterocyst cells are depleted in ^{15}N relative to neighboring vegetative cells, due to high rates of N-export relative to N-fixation. Elevated $\delta^{15}N$ was also observed in intracellular zones within vegetative cells prior to septation; these ^{15}N-rich walls are attributed to newly formed proteins delineating the zone of physical separation of daughter cells. *Trichodesmium* IMS-101 are also capable of fixing both CO_2 and N_2 concurrently throughout the day, yet this species does not contain heterocysts. Using sequentially harvested bacteria, we measured alternating $^{15}N/^{13}C$ temporal enrichment patterns in this species that suggest tightly regulated changes in fixation kinetics. Spatial enrichment features indicate how ^{15}N and ^{13}C "hotspots" are dispersed throughout individual cells, and indicate isolated locations of increased N_2 fixation, sites of amino acid/protein synthesis, and cyanophycin storage granules (Figure 1). Regions of Mo and Fe accumulation suggest heightened N-fixation activity in adjacent groups of cells; both metals are co-factors for the nitrogenase enzyme. This combination of NanoSIMS analysis and high resolution microscopy allows isotopic analysis to be linked to morphological features and holds great promise for fine-scale studies of bacterial metabolism and environmental function.

109

Animal Gene Regulatory Networks

R. Andrew Cameron* (acameron@caltech.edu) and **Eric H. Davidson**

California Institute of Technology, Pasadena, CA

Recent progress in this Genomes to Life Project falls in three areas: (1), Authenticating the sea urchin endomesoderm gene regulatory network at the DNA sequence level; (2), Extending the network to the whole embryo; (3) Advances in gene regulatory network theory. Following is a brief discussion of each area.

Authentication of the Network

The endomesoderm network is now the most extensive and experimentally the best supported of all gene regulatory networks for developmental processes. These are a special class of networks because of the enormous information processing requirements of development; in the history of life they were the last to evolve, and they underlie the extremely complex process of spatial (as well as temporal) patterning of gene expression. Because the endomesoderm network has become the paradigmatic example, and its progress represents the growing edge of the field, it is essential to be able to base its predictions and regulatory logic interactions directly in the genomic regulatory code. This means recovering and subjecting to experimental analysis at the DNA sequence level the *cis*-regulatory modules at the key nodes of the network, the process we term "authenticating the network" (Davidson and Levine, 2005). An imposingly difficult task on the face of it, due to the advantages of the sea urchin system for high throughput *cis*-regulatory research, and the technologies developed in this lab for such research, it has become feasible to carry out cis-regulatory tests across the whole network. At present in this lab there are *cis*-regulatory authentication projects supported by this Project underway on the following genes at major network nodes (see Figure): ***gatae***, ***tbrain***, ***foxa***, ***brn1/2/4***, ***cyclofilin***, ***brachyury***, ***blimp1/krox***; of these, ***cyclofilin*** is complete and a report is In Press. Studies were published last year on ***otx*** (**Yuh et al, 2004**); on ***delta*** (**Revilla and Davidson, 2004**) and on ***wnt8*** (**Minokawa et al, 2005**). In addition there were published functional studies on a new gene added into the network, ***brn1/2/4*** (**Yuh et al, 2005**), and on ***gatae*** (**Lee and Davidson, 2005**). The

* Presenting author

key circuitry of the endomesoderm network is increasingly able to be expressed in the “hardwired” terms of the A’s, C’s, G’s, and T’s of the genomic *cis*-regulatory sequence. It is encouraging that current evidence, published and otherwise, is so far almost entirely corroborating the network linkages that were predicted by the perturbation analyses on the basis of which the network was constructed.

Extending the network to the rest of the embryo

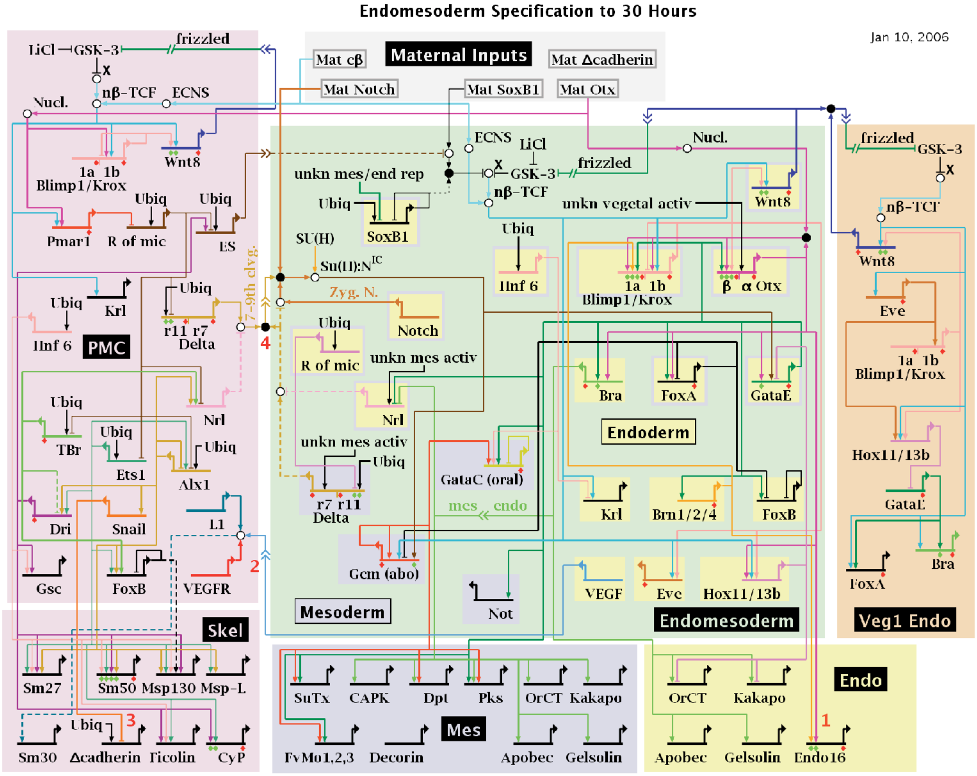
This initiative, one of our original ambitions in this Project, has now been launched. The tremendous reward that will be there for the taking, if we are successful, is that when the components of the whole embryo are included in a network analysis, the overall network will constitute a closed and complete system; after the earliest stages the inputs will all arise within the network. Such a closed, whole organism developmental system has never before been available for study at this level. The effort that has begun is aimed at solving the network for the whole embryo for the same time frame to which the endomesoderm network pertains, ~6-30 hrs after fertilization. The missing parts which are our objective are the oral and aboral ectoderm networks, and the neurogenic apical plate network. Several other labs have begun to work on the latter, and we hope they will be successful, so we won’t have to do it also. The oral and aboral ectoderm networks are a joint project of the laboratory of our subcontractor in the GTL project, David R McClay at Duke, and ourselves. We have devised a new high throughput methodology to build these gene regulatory networks: (1) Determine all regulatory genes that encode sequence specific DNA binding factors predicted in the genomic sequence of this animal, measure their time course of expression, and for those which are expressed by 30 hrs, determine where they are expressed. This is all complete, except for some Zn Finger transcription factor genes, and they will be done soon. (2) For those expressed specifically in oral or aboral ectoderm or both, obtain and verify morpholino substituted antisense oligonucleotides; (3) Carry out a multiplexed perturbation analysis such that the effect of blocking translation of each gene on the expression of all other specific oral and aboral ectoderm regulatory genes is determined at once in a matrix analysis; (4) Utilize a very high density whole transcriptome genomic tiling array chip for the perturbation analysis. (5) Apply the soon to be released version of BioTapestry network building software, which has been built essentially for this purpose, to deconvolve the temporal, spatial, and perturbation data and generate the allowable network architectures. If this new approach works it will revolutionize the practice experimental network building. The network structure will be computationally determined; all expressed genes or any desired subset can be included so the issue of completeness will disappear; and the observations can be multiplexed so that the solution of the network will be vastly accelerated.

Network Theory

During this year the PI worked out a coherent body of new theory for gene regulatory networks which control spatial and temporal gene expression, and which function essentially to set up regulatory states of domains. This will appear in a monograph to be published next year by Elsevier, “**The Regulatory Genome: Gene Networks in Development and Evolution**”. Developmental gene regulatory networks are here treated as networks of information processing nodes, i.e., the individual *cis*-regulatory modules of the network. The ground was laid in previous work undertaken as part of the GTL Project, in which the logic operations integrated within various *cis*-regulatory modules are treated explicitly (**Istrail and Davidson, 2005**). Information processing also emerges from the operations of the subcircuits of which the networks are composed. It is the architecture of these subcircuits which determines the biological functions of the developmental process. The network as a whole is thus to be considered as a large, genomically encoded, delocalized computational device which interprets regulatory information in order to program the dynamic process of development. This “computer” has the capacity to respond conditionally to every possible regulatory state the genome will encounter in each cell of the organism throughout the life cycle.

* Presenting author

Figure 1. The Sea Urchin Endomesoderm Gene Regulatory Network



Ubiq = ubiquitous; Mat = maternal; activ = activator; rep = repressor;
unkn = unknown; Nucl. = nuclearization; χ = β-catenin source;
nβ-TCF = nuclearized b-β-catenin-Tcf1; ES = early signal;
ECNS = early cytoplasmic nuclearization system; Zyg. N. = zygotic Notch

Copyright © 2001-2006 Hamid Bolouri and Eric Davidson

* Presenting author

Section 3

Regulatory Processes

110

VIMSS Computational Core: Comparative Analysis of Regulatory Systems in Environmental Microbes

Eric J. Alm[1,2], Inna Dubchak[1,2], Alex Gilman[1,2], Katherine Huang[1,2], Keith Keller[1,2], and **Adam P. Arkin**[1,2,3]* (APArkin@lbl.gov)

[1]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov; [2]Lawrence Berkeley National Laboratory, Berkeley, CA; and [3]University of California, Berkeley, CA

Background. The VIMSS Computational Core group is tasked with data management, statistical data analysis, modeling, and comparative and evolutionary genomics for the larger VIMSS effort. We have matured many of our analyses and VIMSS data into our flagship comparative functional microbial genomics tool MicrobesOnline (http://microbesonline.org). We have used this framework to interpret the data from the VIMSS physiological pipeline and in more global analysis of genome evolution and function.

Data Analysis. During the course of analysis of the various stress responses of DvH the computational core has developed a number of new statistical analyses of data that take advantage of the predicted regulatory structures (operons, regulons, etc.) from our comparative analyses. We have used these analyses this year to uncover unique responses to salt and heat shock, to exposure to heavy metals, and to elucidate how reducing power is moved among different pathways under nitrogen oxide stress. Both the new methods and the individual analyses have been published recently. We will report on the key finding of these studies.

Data Management. The Experimental Data Repository (http://vimss.lbl.gov/~jsjacobsen/cgi-bin/GTL/VIMSS/datarepository.cgi) has continued to grow in size and functionality and provides access to biomass production data, other growth curve data, synchrotron FTIR data, image data, phenotype microarray data, and transcriptome, proteome and metabolome data. Many of these data types now are supported by sophisticated analyses and visualizations that provide feedback and value to the project personnel and the wider public. This year we will be improving site navigation, project management tools and the range of analyses and visualizations available. It will also be expanded to accept data from collaborating projects in environmental sequences, plant/microbe mesococsms, and protein complex data. Further, a more automated transfer of data from the EDR to MicrobesOnline will be deployed. Many of the data management tools developed as part of this project have be translated to our collaborating projects both related to Genomics:GTL and not.

The MicrobesOnline Database. The MicrobesOnline database (http://microbesonline.org) currently hosts 243 genomes and features a full suite of software tools for browsing and comparing microbial genomes. Highlights include operon and regulon predictions, a multi-species genome browser, a multi-species Gene Ontology browser, a comparative KEGG metabolic pathway viewer and the VIMSS Bioinformatics Workbench for more in-depth sequence analysis. The Workbench provides gene carts that store user defined sets of genes found by searching MicrobesOnline and

* Presenting author

provides tools for multiple sequence alignments, phylogenetic trees construction, and, in prototype, cis-regulatory site detection. We are also incorporating and updating Mikhail Gelfand's highly curated RegTransDb of experimentally verified regulator site/transcription factor data for display on the gene pages of MicrobesOnline and to aid the Workbench cis-regulatory tools.

In addition, we provide an interface for genome annotation, which like all of the tools reported here, is freely available to the scientific community. To keep up with the ever-increasing rate at which microbial genomes are being sequenced, we have established an automated genome import pipeline. A number of outside groups are currently using the MicrobesOnline database for genome annotation projects. To facilitate the use of this community resource we have developed an access control system, so individual research groups can use the power of the VIMSS annotation tools, while keeping data from their own particular genome project private until their analyses are ready to be made public.

Also incorporated in this framework is our microbial microarray analysis suite including a number of quality control metrics, COG/TIGRFAM functional enrichment analysis, operon co-expression analysis and statistical significance tests developed based on this information. The microarray database currently holds data from nine bacteria and twenty-one conditions and it is growing rapidly. The Workbench will be expanded to allow microarray and comparative analysis to be combined. Functional genomics data of all sorts will be directly linked and visualized on the gene pages in MicrobesOnline for which such data exists. As data from our collaborating projects on protein complexes become available this data, too, will be served from this site. We welcome depositing of other peoples' data on MicrobesOnline which in turn we hope will allow users to better analyze their information.

Evolution of Microbial Genomes. We have also used the MicrobesOnline framework to support research on the evolution of specific pathways and genomic architecture in bacteria. We worked with Dmitry Rodionov and Mikhail Gelfand to reconstruct the metabolism of nitrogen oxides in diverse bacteria and later used this information to interpret our functional genomic data on nitrogen oxide stress response in DvH.

We have also used the MicrobesOnline framework for discovering core functionality. For example, we have been able to define a set of "signature" genes that are found in sulfate reducers as diverse of delta-proteobacteria and archaea. We can show that these genes, predicted entirely through comparative sequence analysis, also show coordinated gene expression patterns in DvH.

Additionally, we have derived a novel theory of the full life-cycle of operons in bacteria: how they are born, are tuned, and die. Our findings suggest that operon evolution is driven by selection on gene expression patterns. First, both operon creation and operon destruction lead to large changes in gene expression patterns. For example, the removal of lysA and ruvA from ancestral operons that contained essential genes allowed their expression to respond to lysine levels and DNA damage, respectively. Second, some operons have undergone accelerated evolution, with multiple new genes being added during a brief period. Third, although most operons are closely spaced because of a neutral bias towards deletion and because of selection against large overlaps, highly expressed operons tend to be widely spaced because of regulatory fine-tuning by intervening sequences. Although operon evolution seems to be adaptive, it need not be optimal: new operons often comprise functionally unrelated genes that were already in proximity before the operon formed.

In studying the comparative genomics of the two-component systems that lie at the heart of control of many of the stress responses we are studying we discovered that different organisms use different strategies for generation and acquisition of new sensory histidine kinases. We analyzed the

* Presenting author

phylogenetic distribution of nearly 5000 histidine protein kinases from 207 sequenced prokaryotic genomes. We found that many genomes carry a large repertoire of recently evolved signaling genes, which may reflect selective pressure to adapt to new environmental conditions. Both lineage-specific gene family expansion and horizontal gene transfer play major roles in the introduction of new histidine kinases into genomes; however, there are differences in how these two evolutionary forces act. Genes imported via horizontal transfer are more likely to retain their original functionality as inferred from a similar complement of signaling domains, while gene family expansion accompanied by domain shuffling appears to be a major source of novel genetic diversity. Family expansion is the dominant source of new histidine kinase genes in the genomes most enriched in signaling proteins, such as DvH and other environmental microbes, and detailed analysis reveals that divergence in domain structure and changes in expression patterns are hallmarks of recent expansions. These results lead us to conclude that in the ongoing evolution of bacterial signal transduction machinery, some organisms serve as 'producers' generating novel genetic diversity, while others serve as 'consumers' capitalizing on the existing diversity of their peers.

111

Genome-Wide Mapping of Transcriptional Networks in *Escherichia coli* and *Shewanella oneidensis* MR-1

G. Cottarel, M.E. Driscoll* (mdriscol@bu.edu), J. Faith, M.A. Kohanski, B. Hayete, J. Wierzbowski, C.B. Cantor, **J.J. Collins**, and T.S. Gardner

Boston University, Boston, MA

Both *E. coli* and *Shewanella* possess versatile respiratory networks for energy production, but owing to their distinct ecological niches, this versatility is found at different ends in their redox pathways. *E. coli* thrives on a wide range of electron donors, yet can respire only with oxygen and a few organic compounds. Conversely, *Shewanella*'s use of electron donors is comparatively narrow, but it can respire using dozens of electron acceptors, including solid metals.

The ability to survive in a range of nutritive environments is aided by having the right respiratory enzymes active at the right times. These regulatory responses are conferred by rich transcriptional networks in both *E. coli* and *Shewanella*, which translate changes in environment into changes in gene expression.

To further identify these transcriptional regulatory networks in *E. coli* we have assembled a data set of 505 Affymetrix *E. coli* microarray expression profiles representing over 200 unique conditions and mutants. We assayed 269 of the arrays in our lab, and compiled the other 236 from studies in other laboratories. To our knowledge, this data set represents the largest uniformly collected and normalized microarray dataset for a prokaryote.

We are currently generating a comparable number of expression profiles for *Shewanella*, representing approximately 300 unique growth conditions, as well as mutants generously provided by the *Shewanella* Federation, using an Affymetrix chip we have designed for this organism.

To infer networks from this microarray data, we have developed a novel algorithm based on Bayesian network theory. The algorithm is similar to our NIR algorithm in that it determines the topology of the regulatory network by looking for the most likely regulators of each gene accord-

* Presenting author

ing to gene expression data. Key improvements over the NIR algorithm enable application at the genome-scale and produce more accurate and biologically meaningful regulatory models. These include: (1) the ability to capture nonlinear combinatorial regulatory relationships such as Boolean or thermodynamic gene regulation functions; (2) the ability to incorporate prior information to improve the accuracy of the reverse-engineered network model; and (3) implementation of a Multi-Chain Markov-Chain Monte Carlo numerical optimization algorithm and a model averaging procedure to enable more reliable identification of network topology.

We compared our algorithm's predictions for *E. coli* to RegulonDB, a database of validated transcription factor/promoter interactions. Though RegulonDB is not a complete description of *E. coli* transcription regulation, it is the most comprehensive available. Our algorithm identified 383 regulatory connections, 114 of which have been identified in RegulonDB, and 269 of which are novel.

The wealth of existing knowledge about *E. coli* transcriptional regulation has made it a model organism for improving the performance and validating the results of our network inference algorithms. Though less well-studied, *Shewanella*'s unique metal-reducing capabilities presents an opportunity to explore our transcription network predictions in both bioremediation and microbial fuel cell applications.

112

Computational Approaches to Investigating Transcription Regulatory Networks and Molecular Evolution of the Metal-Reducing Family *Geobacteraceae*

Julia Krushkal[1]* (jkrushka@utmem.edu), Marko Puljic[1], Ronald M. Adkins[1], Jeanette Peeples[1], Bin Yan[1,2], Radhakrishnan Mahadevan[3], and **Derek R. Lovley**[4]

[1]University of Tennessee Health Science Center, Memphis, TN; [2]National Institutes of Health, Bethesda, MD; [3]Genomatica, Inc., San Diego, CA; and [4]University of Massachusetts, Amherst, MA

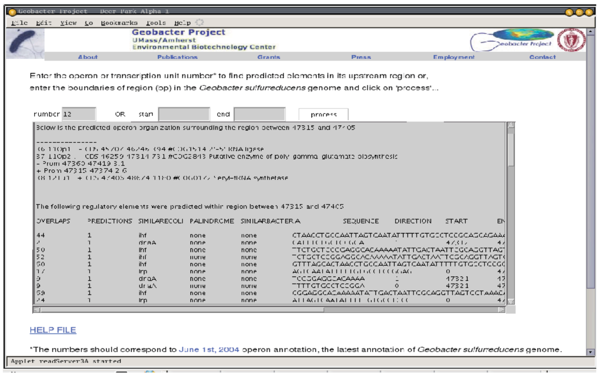
The evolutionary dynamics of gene regulatory mechanisms among living organisms is one of the fundamental questions that is key to our understanding of their ability to adapt to diverse environments and for our ability to manipulate their biology for human needs. We are investigating these mechanisms in *Geobacteraceae*, a metal-reducing family of delta-*Proteobacteria* that participate in bioremediation of contaminated environments and in energy harvesting. We are employing several complementary computational strategies to unveil transcriptional regulatory networks of this environmentally important group of microorganisms and to better understand how genetic differences allow members of this family to adapt to a variety of environmental conditions.

Our transcriptome analysis of delta-*Proteobacteria* involves the prediction of the operon organization of their genomes. We used information obtained from genome sequencing projects of species of delta-*Proteobacteria* to predict operon organization of completed genomes and partial assemblies of *Geobacter sulfurreducens, G. metallireducens, G. uraniumreducens*, *Desulfuromonas acetoxidans* and *D. palmitatis, Pelobacter carbinolicus* and *P. propionicus, Desulfotalea psychrophila, Desulfovibrio desulfuricans, D. vulgaris,* and *Bdellovibrio bacteriovorus.* Operon organization was also predicted for an environmentally important member of beta-*Proteobacteria, Rhodoferax ferrireducens.* The fit of sequence based operon predictions to results of genome wide measurement of gene tran-

* Presenting author

scription and protein production is being evaluated using statistical tests such as ANOVA and permutation analyses.

Figure 1. An example of scored predicted regulatory sites in a user-specified region of the *G. sulfurreducens* genome.



Enter the operon or transcription unit number* to find predicted elements in its upstream region or,
enter the boundaries of region (bp) in the *Geobacter sulfurreducens* genome and click on 'process'...

number 12   OR   start   end

```
Below is the predicted operon organization surrounding the region between 47315 and 47405

----------------
36 110Op1  - CDS 45707 46246 394 #COG1514 2'-5' RNA ligase
37 110Op2 .  CDS 46259 47314 731 #COG2843 Putative enzyme of poly gamma glutamate biosynthesis
- Prom 47360 47419 3.1
+ Prom 47315 47374 2.6
38 121Op1  + CDS 47405 48674 1180 #COG0172 Seryl-tRNA synthetase



The following regulatory elements were predicted within region between 47315 and 47405

OVERLAPS   PREDICTIONS  SIMILARECOLI  PALINDROME  SIMILARBACTERIA      SEQUENCE      DIRECTION    START     EN

44         1            ihf           none        none         CTAACCTGCCAATTAGTCAATATTTTTGTGCCTCCGCAGCAGAA
7          1            dnaA          none        none         CATTTTGTGCCTCCGGA          1             47312     47
50         1            ihf           none        none         TTCTGCTCCCGAGGCACAAAAATATTGACTAATTCGCAGGTTAG
52         1            ihf           none        none         TCTGCTCCGGAGGCACAAAAATATTGACTAATTCGCAGGTTAGT
50         1            ihf           none        none         GTTTAGCACTAACCTGCCAATTAGTCAATATTTTTGTGCCTCCG
17         1            lrp           none        none         AGTCAATATTTTTGTGCCTCCGGAG                   0         47
9          1            dnaA          none        none         TCCGGAGGCACAAAA            1             47321     47
9          1            dnaA          none        none         TTTTGTGCCTCCGGA            0             47321     47
59         1            ihf           none        none         CGGAGGCACAAAAATATTGACTAATTCGCAGGTTAGTCCTAAAC
24         1            lrp           none        none         ATTAGTCAATATTTTTGTGCCTCC                    0         47
```

*The numbers should correspond to June 1st, 2004 operon annotation, the latest annotation of *Geobacter sulfurreducens* genome.

Applet readServer3A started

Building upon our earlier analyses of operon prediction, comparative genomic information, gene expression microarray data, and sequence similarity comparisons, we developed a database containing thousands of predicted transcription regulatory regions of *G. sulfurreducens*. Computational clustering techniques were employed to group those sites that had nearly identical genome locations. Subsequent sorting approaches allowed us to identify those individual sites and regulatory clusters that were predicted by the highest number of gene expression and sequence comparison data sets. These genome-wide predictions pinpoint the regulatory sites and transcription factors that have high likelihood of involvement in important regulatory pathways. For example, the highest scoring regulatory cluster independently predicted from 9 different microarray and sequence prediction sources was located upstream of an operon encoding proteins GSU2941 and GSU2942 that correspond to a transcription regulator from the LysR family and a methyl-accepting chemotaxis protein, respectively. Another high-scoring regulatory cluster predicted from 7 different microarray and sequence analyses sources was located upstream of GSU0655 that corresponds to the σ32 subunit of RNA polymerase (RpoH). These regulators and the upstream regions of their operons likely affect multiple regulatory pathways of *G. sulfurreducens*.

We have developed an online tool that allows a user to query the entire database of predicted regulatory elements in order to identify transcription regulatory sites in a genome region of interest or upstream of an operon of interest (Figure 1). These regulatory elements are ranked based on the number of sources of their prediction. We have utilized this resource to provide ranked predictions of transcription regulatory sites that may regulate expression of dozens of genes important for environmental bioremediation and energy production by *G. sulfurreducens*. Additionally, our database provides thousands of ranked predictions of components of transcription regulatory networks that involve transcription factor-operon interactions. These predicted genome-wide regulatory interactions are now being verified using gene expression microarray data from multiple experiments.

We are using comparative phylogenetic analyses to investigate the molecular evolutionary changes that allowed representatives of *Geobacteraceae* to have a diverse range of physiological and metabolic properties in a variety of environmental conditions. For example, we are investigating the molecular evolution of multiple protein sequences from *Pelobacter carbinolicus,* a cytochrome-poor relative of *G. sulfurreducens*. The pipeline of phylogenetic analyses includes: 1) Blastp sequence similarity searches of microbial proteome data using *P. carbinolicus* and *G. sulfurreducens* protein sequences as queries, 2) ClustalW automated alignments of homologous sequences derived from complete genomes at three levels of phylogenetic divergence: a) other species of delta-*Proteobacteria*, b) all species of *Proteobacteria*, and c) all species of *Bacteria* and *Archea* and 3) calculation of distance matrices and phylogenetic tree inference based on each aligned data set using software packages MEGA, PHYLIP, and PAUP* (Figure 2). In another comparative

* Presenting author

phylogenetic study of transcription regulatory interactions, we are investigating the molecular evolution of each predicted transcription factor of *G. sulfurreducens* by inferring phylogenetic trees containing their homologs from other bacterial and archaeal species.

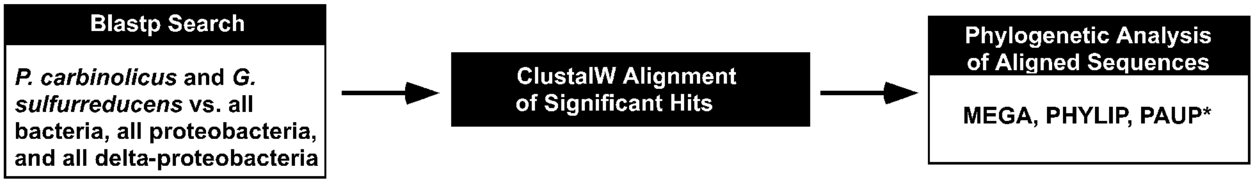

Blastp Search
P. carbinolicus and *G. sulfurreducens* vs. all bacteria, all proteobacteria, and all delta-proteobacteria
ClustalW Alignment of Significant Hits
Phylogenetic Analysis of Aligned Sequences
MEGA, PHYLIP, PAUP*

Figure 2. Pipeline of comparative evolutionary analyses of proteome data of representatives of *Geobacteraceae.*

113

Geobacter Project Subproject IV: Regulatory Networks Controlling Expression of Environmentally Relevant Physiological Responses of *Geobacteraceae*

Byoung-Chan Kim[1]* (bckim@microbio.umass.edu), Katy Juarez-Lopez[1], Hoa Tran[1], Richard Glaven[1], Toshiyuki Ueki[1], Regina O'Neil[1], Ching Leang[1], Gemma Reguera[1], Allen Tsang[1], Robert Weis[1], Marianne Schiffer[2], and **Derek Lovley**[1]

[1]University of Massachusetts, Amherst, MA and [2]Argonne National Laboratory, Argonne, IL

In order to predictively model the physiological response of *Geobacteraceae* under different environmental conditions and to rationally optimize practical applications of these organisms in bioremediation and harvesting electricity from biomass, it is necessary to understand how the expression of genes encoding important physiological functions is regulated. As reported last year, several global regulatory systems, such as RpoS, have now been well characterized in *Geobacter sulfurreducens*. Additional global regulatory systems are being elucidated and more detailed studies on the regulation of key genes are being conducted.

We have recently discovered that *G. sulfurreducens* produces novel, electrically conductive pili that appear to be the electrical conduits to Fe(III) and Mn(IV) oxides and possibly electrodes. The production of these 'microbial nanowires' appears to be highly regulated. For example, at optimal growth temperatures, the nanowires are produced during growth on insoluble Fe(III) or Mn(IV) oxides, but not during growth on soluble electron acceptors. A combination of genetic, microarray, and proteomic investigations revealed that expression of pilA, which encodes the structural protein for the nanowires, is regulated by a two-component regulatory system in which the response regulator, PilR, functions as a sigma 54-EBP (Enhancer Binding Protein). Furthermore, PilR regulates the expression of several c-type cytochromes known to be important in electron transfer to Fe(III). This is the first instance of in which a sigma 54-EBP has been implicated in regulating respiratory functions. In addition, microarray analysis of a mutant incapable of producing acetyl-phosphate suggested that acetyl-phosphate concentrations modulate the activity of PilR, establishing a link between central metabolism and expression of important respiratory genes. These results are important not only for understanding how electron transfer to insoluble electron acceptors is controlled, but also for providing insights into strategies for over-expressing microbial nanowires for practical applications.

* Presenting author

OmcS is an outer membrane c-type cytochrome that is essential for the reduction of Fe(III) oxide and the production of electricity. Expression of omcS was found to be regulated via multiple mechanisms. Evaluation of gene expression in the appropriate mutants revealed that the global regulators Fur, FNR1, and FNR2 all affect the expression of the omcS operon. In addition, omcS expression also appears to be regulated by a two component signal transduction system encoded upstream of the omcS operon. The response regulator associated with this two component system was demonstrated to bind upstream of the omcS start site. We have constructed a knockout mutant lacking the sensor/regulator pair as well as an inducible vector for overexpressing the response regulator in this mutant with the goal of overexpressing the omcS operon for the purposes of optimizing electrical energy harvesting.

The outer-membrane c-type cytochrome OmcB appears to be an essential intermediary electron carrier for extracellular electron transfer. Last year we reported that several outer-membrane c-type cytochromes influence OmcB expression. Some were required for transcription and others were required for translation and/or maturation. In addition, a gene encoding a putative regulatory protease located adjacent to omcF, an outer membrane cytochrome required for omcB transcription, was also found to be required for OmcB production and Fe(III) reduction. These results coupled with recent findings that expression of OmcB is also regulated by RpoS, levels of pppGpp, the transcriptional regulator OrfR, as well PilR demonstrate that multiple regulatory circuits modulate the expression of this central electron transfer component.

Phosphate is often a limiting nutrient in environments in which *Geobacteraceae* predominate. The pho operon, which responds to inorganic phosphate, has been well characterized in other microorganisms, and we have found a homologous two-component signal transduction system that regulates phosphate uptake in *G. sulfurreducens*. We have constructed a knockout mutant lacking the sensor/regulator pair as well as an inducible vector that will overexpress the response regulator. Microarray analysis is being performed to determine the regulatory network that responds to inorganic phosphate availability and to improve regulation inputs for this aspect of the in silico model of *G. sulfurreducens* metabolism. Interestingly, this signal transduction system may act as a global regulator as it influences the transcription of a number of other transcription factors.

The surprising findings that *Geobacter* species are highly planktonic in subsurface environments and that chemotaxis may be an important mechanism for localizing Fe(III) oxides has led to further evaluation of their chemotaxis mechanisms. There are 6 major clusters of chemotaxis genes in *G. sulfurreducens* and 7 in *Geobacter metallireducens*, and these organisms possess large number of chemoreceptors: 20 in *G. metallireducens* and 34 in *G. sulfurreducens*. A novel high throughput signal screening method was developed in order to identify the signals to which these receptors respond. In this assay, individual *Geobacter* chemoreceptors, or chimeric receptors consisting of *Geobacter* sensing domains linked to the cytosolic fragment of the *E. coli* chemoreceptor Tar, are expressed in two *E. coli* strains with defects in chemotaxis, the chemoreceptor-deficient strain UU1250 and the adaptation-deficient strain RP1273, and screened for changes in motility in response to an array of potential signaling molecules. Attractants for a *G. sulfurreducens* chemoreceptor were identified using both the native chemoreceptor and a *Geobacter*/*E. coli* chimera and included acetate and other organic acids. This is significant because acetate is the primary electron donor supporting the growth and activity of *Geobacter* species in subsurface environments and on energy-harvesting electrodes. By fusing signal sensor of histidine kinases to cytosolic fragment of the *E. coli* chemoreceptor, Tar, this strategy can also be used to elucidate the signal specificity of abundant two component systems of the *Geobacteraceae*. We are currently utilizing this approach to identify the environmental signal that the omcS-regulating two component system, described above, responds to.

Additional studies on global regulators including: the sigma factors RpoH, RpoE, and RpoN; the iron response regulators, Fur and IdeR; Fnr; secondary messengers; and other two-component regulatory systems, including novel systems with c-type heme binding motifs in the sensor, are in progress and will be summarized in the presentation.

114

Advances in System Level Analysis of Bacterial Regulation

Lucy Shapiro[1], Michael Laub[2], Ken Downing[3], Alfred Spormann[1], and **Harley McAdams**[1]* (hmcadams@stanford.edu)

[1]Stanford University, Stanford, CA; [2]Harvard University, Cambridge, MA; and [3]Lawrence Berkeley National Laboratory, Berkeley, CA

In the last year we have made major advances in identification of the complete genetic circuitry that runs the cell cycle of the bacterium *Caulobacter crescentus*. We have shown that DnaA acts as a third global regulator that, with CtrA and GcrA, controls the temporal sequencing of the genetic modules that implement the cell cycle. Using custom designed Affymetrix chips, we have identified 769 transcription start sites and 27 conserved promoter motifs used by cell cycle-regulated genes and by genes responding to heavy metal exposure. Two component signal transduction proteins are the primary phosphotransfer system that regulates and coordinates the expressions many cell cycle events. The Laub lab has constructed deletions of each of the 106 *Caulobacter* two component genes and developed a novel technique to identify the targets of the histidine kinases allowing the identification of novel phosphor-regulatory pathways. We are using state-of-the-art fluorescence microscopy and cryo-EM tomography to demonstrate the role of dynamic positioning of regulatory proteins, structural complexes, and chromosomal loci in the overall regulation of the cell as a three dimensional system.

DnaA coordinates replication initiation and cell cycle transcription in *Caulobacter crescentus*. The McAdams and Shapiro labs have shown that DnaA, a critical protein involved in initiation of chromosome replication, also coordinates DNA replication initiation with cell cycle progression by acting as a global transcription factor (Hottes et al. 2005). DnaA functions as a critical transcription factor required for the expression of the gene encoding the GcrA master regulator whose levels oscillate with CtrA during the cell cycle. The redundant control of *gcrA* transcription by DnaA (activation) and CtrA (repression) forms a robust switch controlling the decision to proceed through the cell cycle or to remain in the G1 stage. Other genes in the DnaA regulon include those encoding nucleotide biosynthesis enzymes and components of the DNA replication machinery. Thus, DnaA not only initiates DNA replication, but also promotes the transcription of the components necessary for successful chromosome duplication. DnaA activation of transcription of *ftsZ* and *podJ* starts the cell division and polar organelle development processes that, in addition to DNA replication, prepare the cell for asymmetric division.

Degradation of CtrA requires a dynamically localized ClpXP protease complex and a specificity factor (Iniesta et al.; McGrath et al. 2005). The McAdams and Shapiro labs have shown that the ClpXP protease, which is responsible for the degradation of multiple bacterial proteins, is dynamically localized to specific cellular positions in *Caulobacter,* where it degrades co-localized substrates. For example, the CtrA cell cycle master regulator co-localizes with the ClpXP protease at the stalked cell pole at the swarmer to stalked cell transition and, in the stalked daughter cell compartment

* Presenting author

immediately after cytoplasmic compartmentalization (McGrath, Iniesta et al. 2005), well before daughter cell separation (Judd et al. 2003). C-localization of CtrA with ClpXP is essential for CtrA degradation that enables initiation of chromosome replication. By a combination of bioinformatic, biochemical, genetic, and fluorescent microscopy techniques, we identified two conserved proteins, RcdA and CpdR, that are essential for CtrA degradation. RcdA directly interacts with CtrA and ClpX *in vivo* to mediate CtrA degradation (McGrath, Iniesta et al. 2005). Unphosphorylated CpdR, a response regulator, acts to localize ClpXP at the cell pole (Iniesta, McGrath et al.).

Distinct Constrictive Processes, Separated in Time and Space, Divide *Caulobacter* Inner and Outer Membranes (Judd et al. 2005). Tomographic cryoEM images of the cell division site show separate constrictive processes closing first the inner membrane (IM) and then about 20 minutes later, the outer membrane (OM) in a manner distinctly different from that of septum-forming bacteria. In the early stages of cell division, the inner and outer membranes constrict simultaneously, maintaining their 30-nm separation as seen in regions distant from the constriction. As cell division progresses, the IM constricts faster, creating a growing distance between the inner and outer membranes near the division plane until fission of the inner membrane creates a cell containing two inner membrane-bound cytoplasmic compartments surrounded by a single continuous outer membrane.

High-throughput identification of 769 transcription start sites and 27 DNA regulatory motifs (McGrath et al.). Using 62 data sets of transcription profiles obtained with a custom-designed Affymetrix chip, the McAdams lab identified transcriptional start sites of 769 genes (53 transcribed from multiple start sites. Transcriptional start sites were identified by analyzing the cross-correlation matrices created from the probes tiled every 5 bp upstream of the gene. Motif-searching upstream of the start sites within co-expressed promoters yielded 14 cell cycle regulator binding motifs (8 previously unknown) and 13 heavy metal response regulator motifs (10 previously unknown). This is a ten-fold increase in known *Caulobacter* transcription start sites and a doubling of known binding motifs.

Identification of a Novel Cell Cycle Checkpoint (Spangler et al.).Using DNA microarrays, the Laub lab mapped the response of wild type *Caulobacter* cells to DNA damage and identified two major transcriptional responses: induction of an SOS regulon and repression of genes activated by CtrA, the cell cycle master regulator. Included in the SOS regulon is a novel, but highly conserved gene named *cciA* which is responsible for preventing cell cycle progression after DNA damage. CciA co-localizes with the chromosome segregation machinery component topoisomerase IV and directly inhibits both gyrase and topo IV. These results suggest that DNA damage leads to a delay in cell division by induction of CciA to directly inhibit chromosome segregation which blocks completion of division. The CciA system is the first cell cycle checkpoint identified in *Caulobacter*.

Systematic Analysis of Two-Component Signal Transduction (Skerker et al. 2005). The Laub lab deleted each of the 106 *Caulobacter* two-component signal transduction genes. Thirty-nine of the genes are required for growth, morphogenesis, or cell cycle progression; 9 are essential. A novel technique to identify cognate pairs, called phosphotransfer profiling was developed to identify response regulator targets of histidine kinases. This *in vitro* biochemical technique successfully identifies the specific, *in vivo*-relevant targets of histidine kinases. A library of *Caulobacter* two-component deletion strains was constructed such that each strain harbors a pair of unique bar-codes (20mers) which enable high-throughput fitness analyses to complement microarray and genetic experiments.

Dissection of Genetic Diversity in Subpopulation of *Caulobacter crescentus* Biofilms. The Spormann lab has shown that *C. crescentus* biofilms exhibit a bi-phasic architecture: flat, monolayer biofilms containing interspersed mushroom-like structures that are the result of clonal growth (Entcheva-Dimitrov et al. 2004). Cells grown from the mushroom structures auto-aggregate in liquid

* Presenting author

static or shaken culture, carry increased cell surface hydrophobicity, and show reduced swarming motility. These traits are inheritable. The mushroom phenotype results from a single point mutation in ORF CC3629 (*rfbB*), encoding dTDP-D-glucose-4,6-dehydratase, which is involved in the O-antigen biosynthesis pathway of lipopolysaccharides (LPS), showing that LPS components are critical determinants of *Caulobacter* biofilm architecture. Our results show genetic variants accumulate rapidly in biofilms and represent an important pool and mechanism for generating microbial diversity.

References

1. Entcheva-Dimitrov, P. and A. M. Spormann (2004). "Dynamics and control of biofilms of the oligotrophic bacterium *Caulobacter crescentus*." *J Bacteriol* **186**(24): 8254-66.
2. Hottes, A. K., L. Shapiro and H. H. McAdams (2005). "DnaA coordinates replication initiation and cell cycle transcription in *Caulobacter crescentus*." *Mol Microbiol* **58**(5): 1340-53.
3. Iniesta, A. A., P. T. McGrath, A. Reisenauer, H. H. McAdams and L. Shapiro "A phospho-signal cascade coupled to cytokinesis controls the dynamic positioning and activity of a proteolysis machine critical for bacterial cell cycle progression." In preparation.
4. Judd, E. M., L. R. Comolli, J. C. Chen, K. H. Downing, W. E. Moerner and H. H. McAdams (2005). "Distinct constrictive processes, separated in time and space, divide *Caulobacter* inner and outer membranes." *J Bacteriol* **187**(20): 6874-82.
5. Judd, E. M., K. Ryan, W. E. Moerner, L. Shapiro and H. H. McAdams (2003). "Fluorescence bleaching reveals asymmetric compartment formation prior to cell division in *Caulobacter*." *Proc. Natl. Acad. Aci. USA* **100**(14): 8235-8240.
6. McGrath, P. T., A. A. Iniesta, K. R. Ryan, L. Shapiro and H. H. McAdams (2005). "Controlled degradation of a cell cycle master regulator requires a dynamically localized protease complex and a polar specificity factor." *Cell*: In press.
7. McGrath, P. T., H. Lee, A. A. Iniesta, A. K. Hottes, P. Hu, L. Shapiro and H. H. McAdams "Identification of transcriptional start sites using high density arrays." In preparation.
8. Skerker, J. M., M. S. Prasol, B. S. Perchuk, E. G. Biondi and M. T. Laub (2005). "Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis." *PLoS Biol* **3**(10): e334.
9. Spangler, J. E., M. S. Prasol and M. T. Laub "A Novel Checkpoint System Regulates Cell Cycle Progression after DNA Damage in *Caulobacter crescentus.*" In preparation.

This work was funded in part by the U.S. Department of Energy's Genomics:GTL program under two projects, "Global Characterization of Genetic/Regulatory Circuitry Controlling Adaptive Metabolic Pathways" and "Advanced methods for high-resolution imaging of dynamic cellular and subcellular processes in complex bacterial systems".

115

A System-Level Analysis of Two-Component Signal Transduction

Michael T. Laub[1] (laub@cgr.harvard.edu), Lucy Shapiro[2], and Harley H. McAdams[2]

[1]Harvard University, Cambridge, MA and [2]Stanford University, Stanford, CA

Two-component signal transduction systems, comprised of histidine kinases and their response regulator substrates, are the predominant means by which bacteria sense and respond to signals. These systems allow cells to adapt to prevailing conditions by modifying cellular physiology, including initiating programs of gene expression, catalyzing reactions, or modifying protein-protein interactions. These signaling pathways have also been demonstrated to play a role in coordinating bacterial cell cycle progression and development. We have initiated a system-level investigation of two-component pathways in the tractable model organism *Caulobacter crescentus*, which encodes 62 histidine kinases and 44 response regulators. Comprehensive deletion and overexpression screens have identified more than 40 of these 106 two-component genes as required for growth, viability, or proper cell cycle progression in standard laboratory growth conditions. In addition, the creation of a comprehensive library of deletion mutants enables the identification of two-component signaling genes required for survival in alternative growth conditions and in response to environmental changes. These studies are done in a quantitative and high-throughput manner using a pooled, bar-coding strategy similar to that used for the yeast genome-wide deletion project. Comparison of these experiments with DNA microarray experiments under identical conditions demonstrate that the signal transduction genes *required* for a response show virtually no overlap with the set of signaling genes whose expression level changes.

As with most bacterial species, the majority of genes encoding histidine kinases in *Caulobacter* are orphans – i.e. not encoded in an operon with their cognate response regulator – demanding other approaches for mapping signaling pathways. To address this need we have developed a novel systematic biochemical approach, called phosphotransfer profiling, to map the connectivity of histidine kinases and response regulators. By combining genetic and biochemical approaches, we have begun mapping pathways critical to growth and cell cycle progression. This includes a novel essential two-component signaling pathway, CenK-CenR, which controls cell envelope biogenesis, as well as a complex phosphorelay controlling CtrA, the master regulator of the *Caulobacter* cell cycle. Specific examples will be presented to demonstrate how the combination of techniques, applicable to any bacterial species, can be used to rapidly uncover regulatory networks.

The ability of our *in vitro* phosphotransfer profiling method to identify signaling pathways that are relevant *in vivo* takes advantage of an observation that histidine kinases are endowed with a global, kinetic preference for their cognate response regulators. This system-wide selectivity helps insulate two-component pathways from one another, preventing unwanted cross-talk. Moreover, it suggests that the specificity of two-component signaling pathways is determined almost exclusively at the biochemical level. We are using computational and mutagenesis methods to map the amino acids which confer specificity. The results may enable computational prediction of two-component pairings in any organism and the rational design of novel signaling pathways for constructing biosensors or synthetic genetic circuits.

116

The Bacterial Birth Scar, a Spatial Determinant for the Positioning of a Polar Organelle at the Late Cytokinetic Site

Edgar Huitema, Sean Pritchard, David Matteson, Sunish Kumar Radhakrishnan, and Patrick H. Viollier* (Patrick.Viollier@case.edu)

Case Western University, Cleveland, OH

Many prokaryotic protein complexes underlie polar asymmetry. In *Caulobacter crescentus* a flagellum is built exclusively at the pole that arose from the previous cell division. The basis for this pole-specificity is unclear, but could involve a cytokinetic birth scar that marks the newborn pole as the flagellum assembly site. We identified two novel developmental proteins, MadX and MadR, which localize to the division septum and the newborn pole after division. We show that septal localization of MadX/R depends on cytokinesis. Moreover MadR, a c-di-GMP phosphodiesterase homolog, is a flagellum assembly factor that relies on MadX for proper positioning. In the absence of MadX, flagella are assembled at ectopic locations and MadR is mislocalized to such sites. Thus MadX and MadR establish a link between bacterial cytokinesis and polar asymmetry, demonstrating that division indeed leaves a positional mark in its wake to direct the biogenesis of a polar organelle.

This work was funded in part by the U.S. Department of Energy's Genomics:GTL program under two projects, "*Global Characterization of Genetic/Regulatory Circuitry Controlling Adaptive Metabolic Pathways*" and "*Advanced methods for high-resolution imaging of dynamic cellular and subcellular processes in complex bacterial systems*".

117

Anaerobic Respiration Gene Regulatory Networks in *Shewanella oneidensis* MR-1

Jizhong Zhou[1,2]* (jzhou@ou.edu), Haichun Gao[1,2,3], Xiaohu Wang[5], Yunfeng Yang[2], Xiufeng Wan[6], Zamin Yang[2], Dawn Klingeman[2], Alex Beliaev[4], Soumitra Barua[1,2], Ting Li[2], Christopher Hemme[1,2], Yuri A. Gorby[4], Mary S. Lipton[4], Steve Brown[2], Margaret Romine[4], Kenneth Nealson[7], James M. Tiedje[3], Timothy Palzkill[5], and James K. Fredrickson[4]

[1]University of Oklahoma, Norman, OK; [2]Oak Ridge National Laboratory, Oak Ridge, TN; [3]Michigan State University, East Lansing, MI; [4]Pacific Northwest National Laboratory, Richland, WA; [5]Baylor College of Medicine, Houston, Texas; [6]Miami University, Oxford, OH; and [7]University of Southern California, Los Angeles, CA

Shewanella oneidensis MR-1, a facultative γ-*proteobacterium*, possesses remarkably diverse respiratory capacities. In addition to 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 remain poorly understood. As a part of the *Shewanella* Federation efforts, we have used integrated genomic, proteomic and computational technologies to study the regulatory networks of energy metabolism of this bacterium from a systems-level perspective.

* Presenting author

ArcA. In *Escherichia coli*, metabolic transitions between aerobic and anaerobic growth states occur when cells enter an oxygen-limited condition. Many of these metabolic transitions are controlled at the transcriptional level by the activities of the global regulatory proteins ArcA (aerobic respiration control) and Fnr (fumarate nitrate regulator). A homolog of ArcA (81% amino acid sequence identity) was identified in *S. oneidensis* MR-1, and *arcA* mutants with either MR-1 or Dsp10 (a spontaneous rifampicin-resistant mutant) as the parental strains were generated. The *arcA* deletion mutant grew slower than the wild type and was hypersensitive to H_2O_2. Microarray analysis indicated that S. *oneidensis* ArcA regulates a large number of genes that are not within the *E. coli* ArcA regulon although a small set of genes were found overlapping.

The *S. oneidensis arcA* gene and a mutated *arcA* gene carrying the point mutation of D54N were cloned and expressed in *E. coli*. Both the wild type and the modified ArcA proteins were purified and their DNA binding properties were analyzed by electrophoretic motility shift (EMS) and DNase I footprinting assays. The results indicated that the phosphorylated ArcA proteins were able to bind to a DNA site in the MR-1 genome similar in sequence to the *E. coli* ArcA binding site. The common feature of the binding site is the presence of a conserved 15 bp motif with 2-3 mismatches to the *E. coli* ArcA-P consensus binding motif. Additional EMS experiments revealed that the *S. oneidensis* ArcA proteins can bind to the promoter region of genes whose products are involved in the anaplerotic shunt (*sfcA*), hydrogen metabolism and the terminal DMSO reductase in a phosphorylation-dependent manner. Genome scale computational predictions of binding sites were also performed and 331 putative ArcA regulatory targets were identified. Although the computational screen is in need of refinement, the results suggest the *S. oneidensis* and *E. coli* ArcA-P proteins may differ significantly in terms of the regulation of energy metabolism/respiration. For example, the *ptsG* and *aceBA* are the only two overlapping operons shared between the *E. coli* and *S. oneidensis* ArcA-P modulons. Despite the fact that both ArcA proteins bind to a similar DNA motif, the regulation of aerobic/anaerobic respiration may be more complex than expected in *S. oneidensis*.

EtrA. A homolog of *E. coli Fnr* was identified in MR-1, termed *etrA* (electron transport regulator), which showed a high degree of amino acid identity (51%). An *etrA* deletion mutant was generated but no obvious phenotypic differences under nitrate and fumurate conditions between the wild type and the *etrA* mutant were observed except that the mutant is hypersensitive to H_2O_2. Both strains were individually grown in continuous culture for 410 h under fully aerobic followed microaerobic steady state conditions and then transitioned to anaerobic steady state growth with fumarate as the electron acceptor. Microarray analysis revealed that about 20% of the ORFs showed significant differences in expression between the *etrA* mutant and the parental strains under these growth conditions. The study also revealed that expression profiles of aerobic and microaerobic cultures shared a high level of similarity but greatly differed from that of anaerobic culture. Among genes whose upstream region contains a conserved *E. coli* EtrA-binding motif, only a small number were found significantly affected by the mutation. In addition, the protein expression analysis of the steady state cultures indicated that approximately 20% of all proteins showed significant changes under the steady state conditions. While some genes were significantly changed in both mRNA and protein profiles, significant differences in expression profiles were also noticed. While the exact role of EtrA in *S. oneidensis* remains unknown, it is evident that EtrA of S. *oneidensis* differs from Fnr of *E. coli* significantly in functionalities.

Regulation of nitrate respiration. *S. oneidensis* MR-1 has a NarQ/NarP two-component system. Genome predictions indicate that NarP is the regulator belonging to the LuxR/UhpA family while NarQ is the membrane-anchored sensor. NarP controls the expression of several genes involved in anaerobic respiration and fermentation in *E. coli*. In the presence of nitrate, NarP can be phosphorylated by phospho-NarQ in *E. coli*. In this activated state, phospho-NarP can act as an activator of transcription of the nitrate reductase systems and as a repressor of expression of genes for other anaerobic systems. Two in–frame deletion mutants, *narP* and *narP/narQ*, were generated in MR-1. The *narP* mutant failed to grow on nitrate and nitrate reduction was not evident. In contrast, the *narP/narQ* double mutant grew well on nitrate with much higher biomass yields than MR-1. Nitrate was reduced to ammonia in the *narP/narQ* double mutant whereas nitrate was reduced to nitrite in MR-1. A parallel study on the *S. oneidensis* NAP system demonstrated that NAP was essential for growth of MR-1 on nitrate. These results suggest in *S. oneidensis* that: (1) the NarP/NarQ two-component system controls the NAP system directly; (2) there may be another nitrate reduction pathway through which nitrate can be reduced to ammonia; (3) other regulatory systems controlling the alternative nitrate reduction pathway exist when the NarP/NarQ two-component system is absent. In addition, his-tagged NarP protein has been purified. Results of electrophoretic mobility shift assays (EMSA) showed that phosphorylated NarP was able to bind to its own promoter. The identity of the binding site for the *narQP* operon has been narrowed down to a 74 base pair region in the promoter. Investigations on NarP are focused on defining the binding sites and thereby the NarP regulon of *S. oneidensis*.

Small regulatory RNAs. Small regulatory RNAs have been hailed as the key to coordinate global regulatory circuits. However, essentially nothing is known regarding the potential role of regulatory RNAs in *S. oneidensis*. As an exploratory study, we investigate a small RNA named RyhB, whose counterpart in *E. coli* is known to be regulated by extracellular iron and Fur. RyhB was predicted in *S. oneidensis* and other *Shewanella* species, and then experimentally validated by Reverse Transcription-PCR. Preliminary results suggest that alike *E. coli* RyhB, the expression of *S. oneidensis* homolog is also regulated by iron and Fur. Interestingly, RyhB expression is specifically induced under iron-reducing condition.

In summary, although several global regulatory genes were identified in *S. oneidensis* MR-1 with high degree of amino sequence identity to those in *E. coli*, regulatory circuits in *S. oneidensis* MR-1 appear to have distinct roles in MR-1. This could reflect the complexity of lifestyles of *Shewanella* as well as the diverse environments where the bacterium lives.

* Presenting author

118

Uncovering the Regulatory Network of *Shewanella*

Gary D. Stormo* (stormo@genetics.wustl.edu), Jiajian Liu, and Xing Xu

Washington University School of Medicine, St. Louis, MO

We are working to determine the regulatory sites that control the expression of gene in *Shewanella* species. We are interested in both transcription and post-transcriptional regulation, so we are considering both DNA motifs that bind to transcription factors and RNA motifs that regulate mRNA translation, attenutation and degradation.

Our analysis is being done in two phases. Because much is known about the *E. coli* genome and its regulatory network, we are first determining which factors and binding sites are conserved between the species, allowing us to infer conserved regulatory interactions in those cases. Comparing known and predicted transcription factors (TFs) from *E. coli* to the 5 *Shewanella* species with available genome sequences, we identify 41 TFs that occur in every genome. There are varying numbers that occur in some *Shewanella* species but not all (at least in part this is due to the incompleteness of some of the genome sequences). But for a large number of *E. coli* TFs we find no clear orthologous sequences (based on reciprocal best BLAST matches) in any of the *Shewanella* species. To follow up that study we are also determining which TFs occur only in *Shewanella* and what their distributions are across the different species.

In order to determine the set of conserved intergenic motifs, which represent putative regulatory sites, we are using the program PhyloNet (Wang and Stormo, PNAS 102:17400-5). This program uses profiles from aligned orthologous intergenic regions and compares them across the entire genome to identify those which occur multiple times and are likely to represent regulatory sites for TFs that control multiple genes. Because of the preponderance of dimeric TFs in bacteria, the parameters are the program are being optimized for that type of data. Preliminary results show that most of the regulatory sites for TFs that are conserved with *E. coli* can be identified by this approach. In addition many novel predicted motifs are obtained that probably correspond to TFs found only in the *Shewanella* species.

Our studies of post-transcriptional regulatory sites are initially focusing on identifying RNA structural motifs that are conserved with *E. coli*. Many such post-transcriptional regulatory sites are known in *E. coli*, including several recently described riboswitches. Some of these can be easily identified in *Shewanella*, whereas others are more difficult to detect. We are also beginning the search for novel post-transcriptional regulatory motifs in the 5'UTRs of *Shewanella* genes using software we have developed to identify conserved secondary structures.

* Presenting author

119

Integration of Control Mechanisms and the Enhancement of Carbon Sequestration and Biohydrogen Production by *Rhodopseudomonas palustris*

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

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

The nonsulfur purple (NSP) photosynthetic (PS) bacteria[1] are the most metabolically versatile organisms found on Earth and they have become model organisms to understand the biology of a number of important life processes. One bacterium, *Rhodopseudomonas palustris*, is unique in that it is able to catalyze more processes in a single cell than any other member of this versatile group. Thus, this organism probably catalyzes more fundamentally and environmentally significant metabolic processes than any known living organism on this planet. *R. palustris* is a common soil and water bacterium that can make its living by converting sunlight to cellular energy and by absorbing atmospheric carbon dioxide and converting it to biomass. It is often the most abundant NSP PS bacterium isolated in enrichments. Its abundance is most probably related to one of its unique characteristics; i.e., unlike other NSP PS bacteria, *R. palustris* can degrade and recycle components of the woody tissues of plants (wood contains the most abundant polymers on earth). *R. palustris* can do this both aerobically in the dark and anaerobically in the light. Recent work has shown that regulation of the processes of CO_2 fixation, N_2 fixation, and H_2 metabolism is linked in NSP bacteria[2]. Moreover, a different, yet uncharacterized regulatory mechanism operates under aerobic conditions (unpublished results). Now that its genome sequence is available through the efforts of the JGI and the members of this consortium[3], interactive metabolic regulation of the basic CO_2, hydrogen, nitrogen, aromatic acid, and sulfur pathways of *R. palustris*, as well as other important processes, can be probed at a level of sophistication that was not possible prior to the completion of the genomic sequence. We have pooled the collective expertise of several investigators, using a global approach to ascertain how all these processes are regulated in the cell at any one time. These studies take advantage of the fact that *R. palustris* is phototrophic, can fix nitrogen and evolve copious quantities of hydrogen gas, and is unique in its ability to use such a diversity of substrates for both autotrophic CO_2 fixation (i.e., H_2, H_2S, $S_2O_3^{2-}$, formate) and heterotrophic carbon metabolism (i.e., sugars, dicarboxylic acids, and aromatics, plus many others) under both aerobic and anaerobic conditions.

With regard to the integration and control of basic metabolic processes, we have shown that there is reciprocal regulation of CO_2 fixation and nitrogen fixation/hydrogen metabolism in this organism. Indeed, by blocking the processes by which *R. palustris* removes excess reducing equivalents generated from the oxidation of organic carbon, strains were constructed such that reducing equivalents could be 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 in bio-reactor systems, the advent of these newly isolated strains, in which hydrogen 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

* Presenting author

coupled to the generation of hydrogen gas[4], by combining the properties of the hydrogen-producing derepressed strains, with waste organic carbon degradation, there is much potential to apply these basic molecular manipulations to practical advances. To maximize this capability, the coordinated application of gene expression profiling (transcriptomics), proteomics, carbon flux analysis and bioinformatics approaches have been combined with traditional studies of mutants and physiological/biochemical characterization of cells[5]. During the course of these studies, novel genes and regulators were identified from investigating control of specific processes by conventional molecular biology/biochemical techniques[6]. These studies, along with the microarray and proteomics studies discussed above, have shown that there are key protein regulators that control many different processes in this organism. In many instances, further surprises relative to the role of known regulators, such as the Reg system and CbbR, were noted in *R. palustris*. A novel phospho-transfer system for controlling CO_2 fixation gene expression was also identified and biochemically characterized and a unique signaling process was revealed[7]. Moreover, the key regulator was shown to possess motifs that potentially respond to diverse metabolic and environmental perturbations, suggesting an exquisite means for controlling CO_2 fixation. Likewise, interesting and important genes and proteins that control sulfur oxidation, nitrogen fixation, hydrogen oxidation, and photochemical energy generation have been identified and characterized, and the biochemistry of these systems is under intense study.

In summary, functional analysis of the *R. palustris* proteome and transcriptome, along with traditional biochemical/physiological characterization, has led to considerable progress, placing our group in excellent position to address long term goals of computational modeling of metabolism such that carbon sequestration and hydrogen evolution might be maximized.

References

1. Tabita, F. R., and Hanson, T. E. "Anoxygenic photosynthetic bacteria." 2004. In: *Microbial Genomics*. C. M. Fraser, K. E. Nelson, and T. D. Read (eds.). Humana Press, Inc., Totowa, NJ, pp. 225-243.
2. Dubbs, J. M., and Tabita, F. R. "Regulators of nonsulfur purple phototrophic bacteria and the interactive control of CO_2 assimilation, nitrogen fixation, hydrogen metabolism and energy generation." *FEMS Microbiol. Rev.* **28** (2004) 353-376.
3. Larimer, F.W., Chain, P., Hauser, L., Lamerdin, J., Malfatti, S., Do, L., Land, M., Pelletier, D.A., Beatty, J.T., Lang, A.S., Tabita, F. R., Gibson, J.L., Hanson, T. E., Bobst, C., Torres y Torres, J., Peres, C., Harrison, F.H., Gibson, J., and Harwood, C.S. "Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris*." *Nature Biotechnology* **22** (2004) 55-61.
4. Oda, Y., Samanta, S. K., Rey, F. E., Wu, L., Liu, X., Yan, T., Zhou,J., and C. S. Harwood. "Functional genomic analysis of three nitrogenase isozymes in the photosynthetic bacterium *Rhodopseudomonas palustris*." *J. Bacteriol.* **187** (2005) 7784-7794
5. VerBerkmoes, N. C., M. B. Shah, P.K. Lankford, D. A. Pelletier, M. B. Strader, D. L. Tabb, W. H. McDonald, J. W. Barton, G. B. Hurst, L. Hauser, B. H. Davison, J. T. Beatty, C. S. Harwood, F. R.Tabita, R. L. Hettich, and F. W. Larimer. "Determination and comparison of the baseline proteomes of the versatile microbe *Rhodopseudomonas palustris* under its major metabolic states." *J. Proteome Res.* (in press).
6. Braatsch, S.,J. Bernstein, F. Harrison, J. Morgan, J. Liao, C. Harwood, and J. T. Beatty. "Both of the two *ppsR* genes of *Rhodopseudomonas palustris* CGA009 encode repressors of photosynthesis gene expression." Submitted for publication.
7. Romagnoli, S., and F. R. Tabita. "A novel three-protein two-component system provides a regulatory twist on an established circuit to modulate expression of the *cbb*$_I$ region of *Rhodopseudomonas palustris* CGA010." Submitted for publication.

120

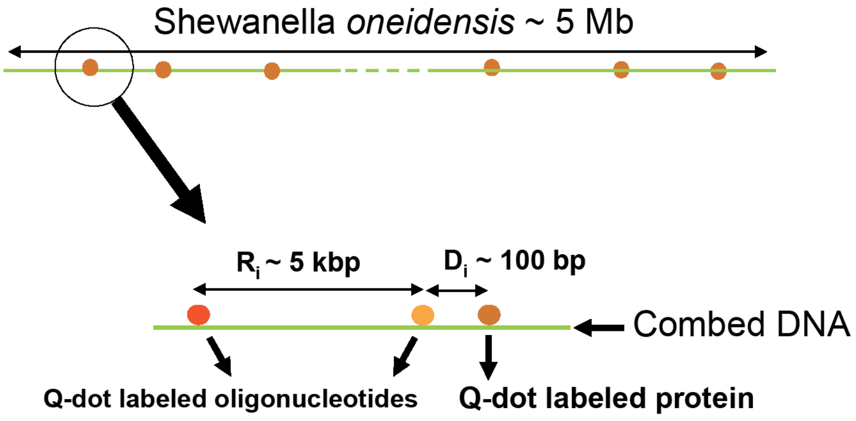
High-Resolution Physical Mapping of Transcription Factor Binding Sites in Whole Bacterial Genomes

J. Antelman[1], X. Michalet[1]* (michalet@chem.ucla.edu), S.B. Reed[2], M. Romine[2], and **S. Weiss**[1]

[1]University of California, Los Angeles, CA and [2]Pacific Northwest National Laboratory, Richland, WA

We propose to develop a novel technique for identifying and mapping regulator protein binding sites on bacterial genomes with very high-resolution, using quantum dots to uniquely tag both DNA and proteins. Our strategy to perform high-resolution physical mapping combines three technologies: DNA molecular combing[1], single-quantum dot labeling and detection[2], and ultrahigh-resolution multicolor colocalization[3]. The dissimilatory metal-reducing bacterium *Shewanella oneidensis* MR-1 will be used as a model system because its genome has been completely sequenced and variety of resources (protein-specific antibodies, purified proteins, clones for expressing and purifying proteins, cloning encoding promoter-encoding DNA) will be accessible from ongoing GTL funded work[4]. We will specifically study the binding sites of the catabolite repressor protein (CRP) protein, for which expression data is already available, but little DNA-protein interaction data. The localization precision (within ~25 bp) of this technique will provide significantly improved data for computational prediction of conserved binding sequences. Furthermore, the positioning of the binding sites relative to the RNA polymerase binding site will be used to predict whether the interaction leads to repression or enhancement of transcriptional activity.

The following diagram summarizes our approach:

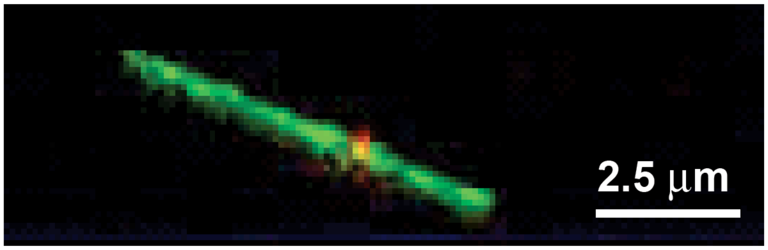

Shewanella *oneidensis* ~ 5 Mb
R_i ~ 5 kbp
D_i ~ 100 bp
Combed DNA
Q-dot labeled oligonucleotides
Q-dot labeled protein

DNA binding proteins bound to DNA *in vivo* are cross-linked prior to DNA extraction. The DNA, with bound protein, is then site-specifically labeled with either a digoxigenin- or a biotin-labeled oligonucleotide probe. Following this, the DNA is stained using a fluorescent intercalating dye and stretched onto a hydrophobic glass surface using DNA molecular combing. After combing, the oligonucleotides are detected with antidigoxigenin-coated or streptavidin-coated quantum dots of various emission wavelength. Additionally, a quantum dot coated with antibodies against the protein of interest is added to visualize the bound protein. Precise measurement of the respective distances between the individual quantum dots (within a few dozen base pair) is then performed using multicolor stage-scanning confocal microscopy as described in ref. [3].

* Presenting author

To identify different oligonucleotide-labeled regions simultaneously, corresponding to different binding sites, unique nearby oligonucleotide probes labeled with digoxigenin and a biotin are used. To distinguish between regions, the probes are separated by different distance (R_1 on the diagram above). Simultaneous AFM imaging is used as a control in the initial stages to ensure that both oligonucleotides and protein are attached to the same combed DNA strand. The end result is a complete, visual map of all binding sites for the studied protein.

Presently, two important accomplishments have been achieved towards the development of this technique. First, a novel method for rendering glass coverslips hydrophobic has been developed. These modified hydrophobic glass coverslips are required for the molecular combing process. The new approach is cheaper, faster and more robust than the previously published technique, and will allow any lab to use this protocol with standard laboratory materials. Secondly, we have modified a RecA mediated oligonucleotide-labeling technique, which allows for the site-specific labeling of dsDNA. These probes can be inserted at any sequence in genomic DNA. Finally, we have demonstrated the detection of this type of probes with single streptavidin- or antidigoxygenin-labeled quantum dots. Shown below is a streptavidin-labeled quantum dot bound to a biotin-labeled probe, hybridized on locus 4404-4434 of lambda phage DNA, counterstained with the DNA intercalating dye YOYO-1.



References

1. Michalet, X., et al., *Science* **277** (1997) 1518
2. Michalet, X., et al., *Science* **307** (2005) 538
3. Lacoste, T., et al., *Proc. Natl. Acad. Sci. USA* **91** (2000) 9461
4. Kolker, E., et al., *Proc. Natl. Acad. Sci. USA* **102** (2005) 2099

* Presenting author

Milestone 3

GTL Computational Biology Environment

Section 1

Computing Infrastructure and Education

121

Research Highlights from the BACTER Program

Julie C. Mitchell* (mitchell@math.wisc.edu), Timothy J. Donohue, George N. Phillips Jr., Nicole Perna, Qiang Cui, David Schwartz, Mark Craven, Paul Milewski, and Stephen Wright

University of Wisconsin, Madison, WI

The BACTER Institute for Computational Biology (http://www.bacter.wisc.edu) has completed its first full year of graduate training activities. We will present vignettes of the research being done by BACTER students and postdocs, in collaboration with the above-mentioned faculty at the University of Wisconsin. Our training and research efforts are directed in three areas: Comparative Genomics, Cellular and Molecular Models, and Biological Pathways. To help focus and interrelate the research of our students, *Rhodobacter sphaeroides* and *Shewanella oneidensis* are the model organisms to which they apply their modeling and informatics toolkits. In comparison with frequent model organisms like yeast or *E. coli*, there is less experimental data available for our model systems. Thus, all of our students, whether in genomics or structural biology, must actively seek out the most recently acquired experimental data and deal with incomplete information. In some cases, our students have initiated their own collaborations with experimental groups to generate the data needed for their research.

Projects in Biological Pathways

We have several students and postdocs working on biological pathways. One student has created a model for the Calvin Cycle in *Rhodobacter*, an organism that has two distinct forms of Rubisco. A BACTER postdoc has performed a highly accurate fitting of experimental data for metal reduction in *Shewanella*, within the context of high-level mathematical models that can be used to predict reaction rates in the presence of multiple metals and metabolites. Two additional students are working on different aspects of microarray data analysis for *Rhodobacter* and *Shewanella*, with the goal of mapping the interactomes of these organisms via machine learning techniques and automated model generation.

Projects in Cellular and Molecular Models

Molecular dynamics and quantum mechanical calculations are being used to predict rate constants across entire biological pathways in *Rhodobacter*. In cases where data is missing, or available only for other organisms, feasible parameter values can fill in the gaps. This undertaking will greatly help advance the research of the Biological Pathways group, and even drive experimental research to help refine and adapt their predictive models. In addition, BACTER is developing accurate protein docking methods able to model large conformational changes with significant dimension reduction in the free variables. Finally, chemotaxis is being studied from the perspective of coupling spatial and

* Presenting author

signaling phenomena into a single mathematical model, which might be neither a partial differential equation nor a system of nonlinear equations, but clearly must have some properties of each.

Projects in Comparative Genomics

We are applying optical mapping technology to *Rhodobacter* genomes, to discover which essential genes are conserved across multiple isolates of its species. The ability to accurately characterize the photosynthetic machinery and carbon sequestration abilities of simple organisms cannot be underestimated. In addition, large-scale comparative genomics tools are being applied to recently sequenced *Shewanella* genomes from JGI, and research into the genetic basis of chemotaxis in *Rhodobacter* will provide nice ties to the research of BACTER's Cellular and Molecular Models group.

122

UC Merced Center for Computational Biology

Michael Colvin[1]* (mcolvin@ucmerced.edu), Arnold Kim[1], Masa Watanabe[1], and Felice Lightstone[2]

[1]University of California, Merced, CA and [2]Lawrence Livermore National Laboratory, Livermore, CA

We have established a Center for Computational Biology (UCM-CCB) at the newest campus of the University of California. The UCM-CCB is sponsoring multidisciplinary scientific projects in which biological understanding is guided by mathematical and computational modeling. The center is also facilitating the development and dissemination of undergraduate and graduate course materials based on the latest research in computational biology. This project is a multi-institutional collaboration including the new University of California campus at Merced, Rice University, Rensselaer Polytechnic Institute, and Lawrence Livermore National Laboratory, as well as individual collaborators at other sites.

The UCM-CCB is sponsoring a number of research projects that emphasize the role of predictive simulations in guiding biological understanding. This research is being performed by post-docs, graduate and undergraduate students and includes mathematical models of cell fate decisions, molecular models of multiprotein machines such as the nuclear pore complex, new mathematical methods for simulating biological processes with incomplete information, and mathematical approaches for simulating the interaction of light with biological materials. The UCM-CCB has run workshops to facilitate computational collaborations with many of the experimental biology programs at UC Merced and is hosting an ongoing seminar series that will bring five prominent computational biologists to speak at UC Merced in Winter and Spring 2006.

Additionally, the UCM-CCB is working to translate this research into educational materials. The UCM-CCB is having a central role in enabling the highly mathematical and computationally intensive Biological Science major, which is currently the largest major at UC Merced and has attracted a very large number of applications for next year. In Fall 2005, the first semester of undergraduate instruction at UC Merced, UCM-CCB materials and expertise were used in computational biology laboratories for two large undergraduate biology courses, and such materials will be used in several more courses in Spring 2006. All course materials are being released under an open public license, and we are in the process of translating these materials into modules in the Connexions courseware system developed at Rice University. The electronic, modular course materials produced by the UCM-CCB are also facilitating linkages to feeder schools at the state university, community college, and high school levels.

* Presenting author

The long-term impact of the CCB will be to help train a new generation of biologists who bridge the gap between the computational and life sciences and to implement a new biology curriculum that can both influence and be adopted by other universities. Such scientists will be critical to the success of new approaches to biology, exemplified by the DOE Genomes to Life program in which comprehensive datasets will be assembled with the goal of enabling predictive modeling of the behavior of microbes and microbial communities, as well as the biochemical components of life, such as multiprotein machines.

123

The BioWarehouse System for Integration of Bioinformatics Databases

Tom Lee, Valerie Wagner, Yannick Pouliot, and **Peter D. Karp*** (pkarp@ai.sri.com)

SRI International, Menlo Park, CA

BioWarehouse[1] is an open-source toolkit for constructing bioinformatics database (DB) warehouses. It allows different users to integrate collections of DBs relevant to the problem at hand. BioWarehouse can integrate multiple public bioinformatics DBs into a common relational DB management system, facilitating a variety of DB integration tasks including comparative analysis and data mining. All data are loaded into a common schema to permit querying within a unified representation.

BioWarehouse currently supports the integration of UniProt, ENZYME, KEGG, BioCyc, NCBI Taxonomy, CMR, Gene Ontology, and the microbial subset of Genbank. Loader tools implemented in the C and Java languages parse and load the preceding DBs into Oracle or MySQL instances of BioWarehouse.

The BioWarehouse schema supports the following bioinformatics datatypes: chemical compounds, biochemical reactions, metabolic pathways, proteins, genes, nucleic acid sequences, features on protein and nucleic-acid sequences, organism taxonomies, and controlled vocabularies.

BioWarehouse is in use by several bioinformatics projects. An SRI project is developing algorithms for predicting which genes within a sequenced genome code for missing enzymes within metabolic pathways predicted for that genome[2]. BioWarehouse fills several roles within that project: it is used to construct a complete and nonredundant dataset of sequenced enzymes by combining protein sequences from the UniProt and PIR DBs, and by removing from the resulting dataset those sequences that share a specified level of sequence similarity. Our current research involves extending the pathway hole filling algorithm with information from genome-context methods such as phylogenetic signatures, which are obtained from BioWarehouse thanks to the large all-against-all BLAST results stored within CMR. Another SRI project is comparing the data content of the EcoCyc and KEGG DBs using BioWarehouse to access the KEGG data in a computable form.

BioWarehouse is supported by the Department of Energy and by DARPA through the DARPA BioSPICE program for biological simulation.

References

1. BioWarehouse Home Page http://bioinformatics.ai.sri.com/biowarehouse/
2. Green, M.L. and Karp, P.D., "A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases," *BMC Bioinformatics* 5(1):76 2004 http://www.biomedcentral.com/1471-2105/5/76.

* Presenting author

124

Communicating Genomics:GTL

Anne E. Adamson, Shirley H. Andrews, 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)

Oak Ridge National Laboratory, Oak Ridge, TN

Project Web Site: DOEGenomesToLife.org

To help build the critical multidisciplinary community needed to advance systems microbiology research, the Genome Management Information System (GMIS) contributes to DOE Genomics: GTL program strategies. GMIS also communicates key scientific and technical concepts emanating from GTL and related programs to the scientific community and the public. We welcome ideas for extending and improving communications and program integration to represent GTL science more comprehensively.

Accelerating GTL Science

For the past 17 years, we have focused on presenting Human Genome Project (HGP) information and on imparting knowledge to a wide variety of audiences. Our goal always has been to help ensure that investigators could participate in and reap the genomic revolution's scientific bounty, new generations of students could be trained, and the public could make informed decisions regarding complicated genetics issues. Since 2000, GMIS has built on this experience to communicate about the DOE Office of Science's Genomics:GTL program.

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 communication approach. Because each discipline has its own perspective, effective communication 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 results reporting 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 its 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. A new research community centered around the advanced concepts in GTL will emerge from this effort.
- Help reduce duplication of effort.

* Presenting author

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

During the past 2 years, we have worked 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, is being reviewed by the National Academy of Sciences. Related tasks have included helping to organize workshops, capture workshop output, and conduct the myriad activities associated with creating a technical document of the roadmap's size and importance. In the past 5 months, as part of the mission goals set forth in the GTL roadmap, we have helped to organize the Biomass to Biofuels workshop and produce the draft report of a joint GTL–DOE Energy Efficiency and Renewable Energy research plan for lignocellulosic bioethanol.

For outreach and to increase program input and grantee base, we identify venues for special GTL symposia and 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 Society for Industrial Microbiology, Plant and Animal Genomes, American Society of Plant Biologists, American Chemical Society, National Science Teachers Association, National Association for Biology Teachers, and Biotechnology Industry Organization.

We mail some 1600 packages of technical and educational material each month to requestors and furnish handouts in bulk to meeting organizers who are hosting genomics educational events. We continue to create and update handouts, including a primer that explores the impact of genomics on science and society and flyers on careers in genetics and other 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.

All GTL publications are on the public web site, which includes an image gallery, research abstracts, and links to program funding announcements and individual researcher web sites. Additional site enhancements are being developed and implemented, including specific pages for DOE missions that will be impacted by the GTL program and its facilities. 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, our 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.

This research is sponsored by the DOE Office of Biological and Environmental Research and Office of Advanced Scientific Computing Research.

* Presenting author

Ethical, Legal, and Societal Issues

125

The DNA Files®

Bari Scott* (bariscot@aol.com)

SoundVision Productions®, Berkeley, CA

SoundVision's highly acclaimed series *The DNA Files®* has twice demonstrated that complex science could be made clear and exciting to listeners with no science background. *The DNA Files 3* will continue to show the general public the importance of cutting-edge science to their everyday lives and, at the same time, expand into a larger audience of minority and rural listeners. In addition, we plan to extend the impact of *The DNA Files 3* beyond the airwaves into schools, museums, news outlets, and even the corner coffee shop through a new network of outreach services, media projects, and learning programs along with our rich and informative web site.

The DNA Files 3 will include five nationally distributed, hour-long public radio documentaries and five 5-minute features exploring the revolutionary new developments in systems biology, neurobiology, immunology and the interaction of the environment with our DNA. The programs tackle complex topics in a style that is evocative, creative and accessible. Scientists, government officials and corporate spokespeople offer their points of view ‹ as do individuals who have direct personal experience with the world of genetic research. Program topics will include: Ethics Beyond the Genome: Systems Biology and Nanotechnology; Beyond DNA: The RNA World and Immunology; Individualizing the Genome: Toxicogenomics and Pharmacogenomics; Our Common Genes: Bugs, Mice and the Human Body; Depression, Addiction and Our Genes: Neurogenetics. In addition to the documentaries themselves, SoundVision has added components to *The DNA Files®* project that will promote science journalism by less-experienced science journalists and in the ethnic media; engage the public directly in workshops employing some of the science concepts covered; and contribute to science education and journalism across a broad spectrum of platforms. The project's new outreach and education services, many of which are geared towards increasing minority and rural audiences, include:

- **Media support and training**

With the goal of expanding and enhancing coverage that makes science clearly relevant to a diverse population, *The DNA Files 3* will make a variety of resources available for journalists and media outlets. This and related outreach will extend the documentaries' impact by encouraging local and regional initiatives and motivating additional programs and news stories outside of SoundVision.

Materials and resources: SoundVision will provide talk show discussion topics targeted to specific ethnic communities as well as general audiences; lists of experts whom reporters can use as sources for programs and follow-up news reports, and highlights from our five documentaries. We will also develop *The DNA Files® Style Book,* a handbook of best practices in science reporting on genetics and molecular biology geared to both minority-owned media and public radio as a whole. In a "push" element, SoundVision will send out short news alerts to reporters and editors to identify news stories related to *The DNA Files 3* documentaries. And, working with two ethnic media consortia, we will provide targeted background material for ethnic print.

* Presenting author

Funding and support: In addition to making the above resources generally available, SoundVision will work directly with up to 20 public radio stations. Technical support and grants (from another funding source) will be made available to stations on a competitive basis for local programming and projects related to *The DNA Files 3.* Stations will be encouraged to develop community outreach strategies and to produce programs ranging from feature reports and local documentaries to call-in programs. The impact is likely to be significant: If each of the 20 stations awarded outreach grants produced just one hour of programming, that alone would quadruple the amount of on-air content resulting from *The DNA Files 3* funding. Our outreach consultants will help these stations develop community interest, build partnerships with ethnic and minority press, and work with local science and informal learning centers.

- **Educational programs**

The Exploratorium, the world-renowned museum of science, art, and human perception in San Francisco, has demonstrated its support for *The DNA Files®* by agreeing to create up to five hands-on workshops that can be produced at the Exploratorium and other science museums around the country. The Exploratorium also will create a *DNA Files Workshop Kit* with materials to provide hands-on learning experiences. The kit will be designed for use by other museums, schools, churches, parks and recreation departments, as well as parents who home-school their children. It will be distributed to these outlets in tandem with the airing of *The DNA Files 3* documentaries.

- **Enhanced website**

Our information-packed multimedia website will provide toolkits to help reporters, editors, museum directors, teachers, and home-schooling parents build articles and lesson plans around *The DNA Files 3* programs. Materials on the site will include original in-depth articles related to each of the five documentaries, background information and research for editors and reporters, a library of links to related web sites, and *The DNA Files® Style Book. The DNA Files 3's* improved website will support public radio programming and museum and school programs that can stimulate public interest in science long after the series airs.

Evaluation

An independent firm specializing in educational media will evaluate *The DNA Files 3.* Its principals will conduct annual online user surveys and interview listeners to gage their understanding and retention of the project's core themes. The evaluation includes an assessment of our relationship with outreach stations, plus pre- and post-production focus groups with multicultural audiences, African American audiences, and high school and junior college level biology teachers. The evaluators concluded that regarding informal educational outcomes that "the style and format were highly effective in raising comprehension and awareness of the content among the focus group participants." They further state that the "producers of *The DNA Files®* have established an effective, appealing model for blending traditional and nontraditional public radio science formats with valuable awareness-building content."

Project History

SoundVision Productions®, a 501(c)(3) nonprofit organization, has produced two previous *The DNA Files®* radio projects and a related multimedia website supported by major grants from a variety of sources. *The DNA Files®* has won numerous awards, including the George Foster Peabody Award, the Alfred I. duPont-Columbia University Award, the American Association for the Advancement of Science Journalism Award, the Robert Wood Johnson Foundation Award, the Society of Professional Journalists Excellence in Journalism Public Service Broadcast Award, the American

* Presenting author

Institute of Biological Sciences Media Award for Broadcast Journalism, and the Association of Women in Communications Clarion Award. National Public Radio, which aired *The DNA Files®*, will continue to air *The DNA Files 3,* which is hosted by John Hockenberry and guided by an outstanding panel of advisors, to its member stations.

126

Science Literacy Project for Public Radio Journalists

Bari Scott* (bariscot@aol.com)

SoundVision Productions®, Berkeley, California

In the genomic era, journalists bear a greater responsibility than ever to communicate science's rapid advances and their societal implications to the public. Understanding the basics about 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 clear and accurate information to the public.

SoundVision Productions® presents three new workshops in its lauded Science Literacy Project series: October 16–22, 2005, in Boston, co-hosted by WGBH-FM and The Whitehead Institute; March 2006 in San Francisco at KQED-FM; and October 2006 in Austin, Texas, in cooperation with KUT-FM at the University of Texas.

In each intensive six-day training workshop, twelve competitively selected, mid-career public radio journalists gain the tools and knowledge to tackle complex science stories in a perceptive, clear and imaginative way.

"I am a better reporter because of what I learned at that workshop, [and] I use what I learned there almost every day."

Today's scientific developments affect all of us. In this technologically advanced and rapidly changing world, the general public needs to grasp not only the science itself, but also its interaction with economics, politics and public policy.

Yet public radio journalists face tremendous challenges as they strive to present such complex information to their audiences. Their hurdles include keeping up with fast-paced science and identifying reliable sources quickly. The creative challenges are also immense: how to unfold a multilayered story using only sound.

The Science Literacy Project addresses these issues.

The SoundVision workshops incorporate three goals. They are designed to increase the number and the quality of science stories produced for radio; increase the number of reporters able to report competently on complex research processes, discoveries, and resulting societal implications; and ultimately, increase civic literacy and help minimize the widening knowledge gap between the scientific community and the public. Workshop lectures explore the interactions of DNA, RNA, and proteins and the overall complexity of the machinery of the cell; teach reporters what scientists are discovering about the most basic elements of life; examine the characteristics of one or more nonpathogenic

* Presenting author

microbes of environmental importance; and delve into the interactions between the human genome and the environment and toxicogenomics. A key workshop focus is ethics in the post-human-genome-project era, including new questions about the relationship between science and business, the impact of highly patented science on society, and the risks and responsibilities of attempting to manipulate life.

Each workshop centers around 15 to 20 presentations by scientists, science journalists, scientific researchers, and radio production professionals. Sessions orient producers to basic science, focus on the craft and responsibility of science journalism, and explore techniques for presenting complex scientific content on radio. The focus on radio is particularly important given the specific production needs that distinguish radio from other media. The trainees will develop additional insight from constructive critiques of their work, support materials, and two follow-up teleconferences. Each workshop will include a field trip and several informal gatherings with scientists to develop relationships and learn more about their ideas and research. The project also includes a website that provides transcripts and selected audio from the training sessions, "tip sheets" and online resources for participants and interested users. Follow-up teleconferences will support participants in pursuing complex and rewarding science stories for their communities.

By the end of the six days, the journalists will have sharpened their capacity to probe into emerging scientific issues, to transform their findings into compelling radio, and to enjoy greater confidence in their science reporting abilities.

"As a liberal arts major I confess I was worried I'd have a hard time getting my head around physics, statistics, DNA, gene splicing and biotechnology. But your presenters made it all very understandable, and dare I say it, fun."

SoundVision's training methodology has had long-lasting positive effects on public radio journalists who have attended previous workshops. Even years after attending, participants from rural to large metropolitan stations report that the workshops still help them with 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 better skills in getting behind press releases and scientific papers to create compelling public radio features. SoundVision's innovative workshops boost participants' confidence and their ability to communicate complex and emerging scientific research accurately. We believe that their collective work throughout the country will help lead to a better public understanding of current scientific research and its social implications.

As in our previous workshop projects, a comprehensive evaluation is conducted by Rockman et al, a well-established, San Francisco-based evaluation firm with expertise in evaluating media projects and assessing the impact of training on journalistic practice.

"I left the workshop feeling both inspired with creative approaches to science radio and armed with a formidable set of tools to help me produce better science programming."

Appendix 1: Participants

As of January 27, 2006

Michael Adams
University of Georgia
adams@bmb.uga.edu

Muktak Aklujkar
University of Massachusetts, Amherst
muktak@microbio.umass.edu

Sergej Aksenov
Gene Network Sciences, Inc
kelly@gnsbiotech.com

Gary Andersen
Lawrence Berkeley National Laboratory
glandersen@lbl.gov

Carl Anderson
Brookhaven National Laboratory
cwa@bnl.gov

Gordon Anderson
William R. Wiley Environmental Molecular Sciences Laboratory
gordon@pnl.gov

Jon Apon
Scripps Research Institute
jvapon@scripps.edu

Michael Appel
Brookhaven National Laboratory
mappel@bnl.gov

Adam Arkin
Lawrence Berkeley National Laboratory
aparkin@lbl.gov

Nacyra Assad-Garcia
J.C. Venter Institute
nassad-garcia@venterintitute.org

Deanna Auberry
Pacific Northwest National Laboratory
deanna.auberry@pnl.gov

Kristin Balder-Froid
Lawrence Berkeley National Laboratory
khbalder-froid@lbl.gov

Nicole Baldwin
Oak Ridge National Laboratory
baldwinne@ornl.gov

Nitin Baliga
Institute for Systems Biology
nbaliga@systemsbiology.org

Robert Balint
CytoDesign, Inc
rfbalint@sbcglobal.net

Jill Banfield
University of California, Berkeley
jill@seismo.berkeley.edu

Ryan Bannen
University of Wisconsin, Madison
rmbannen@wisc.edu

Gang Bao
Georgia Institute of Technology and Emory University
gang.bao@bme.gatech.edu

Soumitra Barua
Oak Ridge National Laboratory
baruas@ornl.gov

John Battista
Louisiana State University
jbattis@lsu.edu

Paul Bayer
U.S. Department of Energy
paul.bayer@science.doe.gov

Alex Beliaev
Pacific Northwest National Laboratory
Imalyn.Knight@pnl.gov

Gwynedd Benders
J. Craig Venter Institute
gbenders@venterinstitute.org

Frank Bergmann
Keck Graduate Institute
fbergman@kgi.edu

Michael Betenbaugh
Johns Hopkins University
beten@jhu.edu

Mark Biggin
Lawrence Berkeley National Laboratory
MDBiggin@lbl.gov

Jeffrey Blanchard
University of Massachusetts
blanchard@microbio.umass.edu

Jennifer Bownas
Oak Ridge National Laboratory
bownasjl@ornl.gov

Timothy J. Boyle
Sandia National Laboratories
tjboyle@sandia.gov

Andrew Bradbury
Los Alamos National Laboratory
amb@lanl.gov

Michael Bramucci
DuPont CR&D
Michael.G.Bramucci@usa.dupont.com

Fred Brockman
Pacific Northwest National Laboratory
fred.brockman@pnl.gov

Michelle Buchanan
Oak Ridge National Laboratory
buchananmv@ornl.gov

Anthony Burgard
Genomatica, Inc.
aburgard@genomatica.com

R. Andrew Cameron
California Institute of Technology
acameron@caltech.edu

Denise Casey
Oak Ridge National Laboratory
caseydk@ornl.gov

Jean Chin
National Institutes of Health
chinj@nigms.nih.gov

Paul Choi
Harvard University
pjchoi@fas.harvard.edu

Linda Chrisey
Office of Naval Research
chrisel@onr.navy.mil

Ray-Yuan Chuang
J. Craig Venter Institute
rchuang@venterinstitute.org

Bruce Church
Gene Network Sciences, Inc.
kelly@gnsbiotech.com

James Cole
Michigan State University
colej@msu.edu

Frank Collart
Argonne National Laboratory
fcollart@anl.gov

Michael Colvin
University of California, Merced
mcolvin@ucmerced.edu

Rita Colwell
University of Maryland
rcolwell@umiacs.umd.edu

Sean Conlan
Wadsworth Center
sconlan@wadsworth.org

Maddalena Coppi
University of Massachusetts, Amherst
mcoppi@microbio.umass.edu

Sarah Cunningham
University of Wisconsin, Madison
sccunningham@wisc.edu

Bruce Dale
Michigan State University
bdale@mail.egr.msu.edu

Sherrika Daniel-Taylor
Sandia National Laboratories
sddanie@sandia.gov

Debopriya Das
Lawrence Berkeley National Laboratory
ddas2@lbl.gov

Madhukar Dasika
Pennsylvania State University
msd179@psu.edu

Brian Davison
Oak Ridge National Laboratory
bod@ornl.gov

Dave DeCaprio
Broad Institute
daved@alum.mit.edu

Vincent Denef
University of California, Berkeley
vdenef@berkeley.edu

Gennady Denisov
J. Craig Venter Institute
gdenisov@venterinstitute.org

Thomas DiChristina
Georgia Institute of Technology
thomas.dichristina@biology.gatech.edu

Alleson Dobson
J. Craig Venter Institute
adobson@venterinstitute.org

Mitchel Doktycz
Oak Ridge National Laboratory
doktyczmj@ornl.gov

Ming Dong
Lawrence Berkeley National Laboratory
mdong@lbl.gov

Timothy Donohue
University of Wisconsin, Madison
tdonohue@bact.wisc.edu

Daniel Drell
U.S. Department of Energy
daniel.drell@science.doe.gov

Inna Dubchak
Lawrence Berkeley National Laboratory
ildubchak@lbl.gov

Roger Ely
Oregon State University
ely@engr.orst.edu

Don Erbach
U.S. Department of Agriculture
dce@ars.usda.gov

Matthew Fields
Miami University
fieldsmw@muohio.edu

Jeffrey Fox
Gene Network Sciences, Inc.
kelly@gnsbiotech.com

Marvin Frazier
J. Craig Venter Institute
mfrazier@venterinstitute.org

Jim Fredrickson
Pacific Northwest National Laboratory
tara.hoyem@pnl.gov

Yuan Gao
Harvard Medical School
yuan888_98@yahoo.com

Roxanne Garland
U.S. Department of Energy
roxanne.garland@ee.doe.gov

George Garrity
Michigan State University
garrity@msu.edu

Natalie Gassman
University of California, Los Angeles
ngassman@chem.ucla.edu

Daniel Gibson
J. Craig Venter Institute
dgibson@venterinstitute.org

Carol Giometti
Argonne National Laboratory
csgiometti@anl.gov

Participants

John Glass
J. Craig Venter Institute
jglass@venterinstitute.org

Robin Goodwin
Michigan State University
goodwi10@msu.edu

Yuri Gorby
Pacific Northwest National Laboratory
Yuri.Gorby@pnl.gov

Stephen Gould
WTEC, Inc.
sgould@nsf.gov

Michelle Green
SRI International
green@ai.sri.com

Masood Hadi
Sandia National Laboratories
MZHADi@sandia.gov

James Hainfeld
Brookhaven National Laboratory
hainfeld@bnl.gov

Bruce Hamilton
National Science Foundation
bhamilto@nsf.gov

Rasha Hammamieh
Walter Reed Army Institute of Research
rasha.hammamieh@na.amedd.army.mil

Terry C. Hazen
Lawrence Berkeley National Laboratory
tchazen@lbl.gov

Zhili He
University of Oklahoma
zhili.he@ou.edu

Grant Heffelfinger
Sandia National Laboratories
gsheffe@sandia.gov

Fred Heineken
National Science Foundation
fheineke@nsf.gov

Bernadette Hernandez-Sanchez
Sandia National Laboratories
baherna@sandia.gov

Robert Hettich
Oak Ridge National Laboratory
hettichrl@ornl.gov

Colin Hill
Gene Network Sciences, Inc
kelly@gnsbiotech.com

David Hill
Dana-Farber Cancer Institute
david_hill@dfci.harvard.edu

Kristina Hillesland
University of Washington
hilleskl@u.washington.edu

Michael Himmel
National Renewable Energy Laboratory
mike_himmel@nrel.gov

Roland F. Hirsch
U.S. Department of Energy
roland.hirsch@science.doe.gov

Dawn Holmes
University of Masschusetts
dholmes@microbio.umass.edu

Norman Hommes
Oregon State University
hommesn@onid.orst.edu

Brian Hooker
Pacific Northwest National Laboratory
brian.hooker@pnl.gov

Tara Hoyem
Pacific Northwest National Laboratory
tara.hoyem@pnl.gov

Greg Hurst
Oak Ridge National Laboratory
hurstgb@ornl.gov

Prabha Iyer
J. Craig Venter Institute
piyer@venterinstitute.org

Chavonda Jacobs-Young
U.S. Department of Agriculture
cjacobs@csrees.usda.gov

Barbara Jasny
Science/AAAS
bjasny@aaas.org

Frank Jenney
University of Georgia
fjenney@bmb.uga.edu

Jian Jin
Lawrence Berkeley National Laboratory
jjin@lbl.gov

Margaret Johnson
National Science Foundation
mdjohnso@nsf.gov

Erik Johnson
University of Chicago
erikj@uchicago.edu

Marc Jones
U.S. Department of Energy
marc.jones@science.doe.gov

Michael Kahn
U.S. Department of Energy
Michael.Kahn@science.doe.gov

Ed Kaleikau
U.S. Department of Agriculture
ekaleikau @csrees.usda.gov

Samuel Kaplan
University of Texas Health Science Center
at Houston
samuel.kaplan@uth.tmc.edu

Peter Karp
SRI International
pkarp@ai.sri.com

Pushpa Kathir
U.S. Department of Agriculture
pkathir@csrees.usda.gov

Arthur Katz
U.S. Department of Energy
arthur.katz@science.doe.gov

Martin Keller
Diversa Corporation
mkeller@diversa.com

Vladimir Kery
Pacific Northwest National Laboratory
vladimir.kery@pnl.gov

William Kimmerly
Pacific Northwest National Laboratory
bill.kimmerly@pnl.gov

Barbara Knutson
University of Kentucky
bknutson@engr.uky.ledu

Eugene Kolker
BIATECH Institute
ekolker@biatech.org

Julia Krushkal
University of Tennessee Health Science Center
jkrushka@utmem.edu

Mike Kuperberg
U.S. Department of Energy
Michael.Kuperberg@science.doe.gov

Cheryl Kuske
Los Alamos National Laboratory
kuske@lanl.gov

Raman Lall
BACTER Institute
lall@wisc.edu

Timothy Lambert
Sandia National Laboratories
tnlambe@sandia.gov

Miriam Land
Oak Ridge National Laboratory
landml@ornl.gov

Elizabeth Landorf
Argonne National Laboratory
elandorf@anl.gov

Carolyn Larabell
Lawrence Berkeley National Laboratory
calarabell@lbl.gov

Carole Lartigue
J. Craig Venter Institute
clartigue@venterinstitute.org

Jared Leadbetter
California Institute of Technology
jleadbetter@caltech.edu

Adam Lee
University of Maryland
adamlee@umd.edu

Dmitriy Leyfer
Gene Network Sciences, Inc
kelly@gnsbiotech.com

Huilin Li
Brookhaven National Laboratory
hli@bnl.gov

Ann Lichens-Park
U.S. Department of Agriculture
apark@csrees.usda.gov

Xiaoxia Lin
Harvard Medical School
xiaoxia@genetics.med.harvard.edu

Liang-Shiou Lin
U.S. Department of Agriculture
llin@csrees.usda.gov

Jan Liphardt
University of California, Berkeley
liphardt@physics.berkeley.edu

Mary Lipton
Pacific Northwest National Laboratory
mary.lipton@pnl.gov

Chang-Jun Liu
Brookhaven National Laboratory
cliu@bnl.gov

Kenton Lohman
U.S. Department of Energy
kenton.lohman@science.doe.gov

Yuri Londer
Argonne National Laboratory
londer@anl.gov

Derek Lovley
University of Massachusetts
dlovley@microbio.umass.edu

Ronald Mackenzie
University of Texas Health Science Center
at Houston
Ronald.C.Mackenzie@uth.tmc.edu

Lee Makowski
Argonne National Laboratory
lmakowski@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

Rodger Martin
U.S. Army Center for Environmental
Health Research
rodger.k.martin@us.army.mil

Anthony Martino
Sandia National Laboratories
martino@sandia.gov

Uljana Mayer
Pacific Northwest National Laboratory
uljana.mayer@pnl.gov

Harley McAdams
Stanford University
hmcadams@stanford.edu

Lee Ann McCue
Pacific Northwest National Laboratory
leeann.mccue@pnl.gov

Gerry McDermott
Lawrence Berkeley National Laboratory
gmcdermott@lbl.gov

Gail McLean
U.S. Department of Agriculture
gmclean@csrees.usda.gov

Angeli Menon
University of Georgia
almenon@uga.edu

Noelle Metting
U.S. Department of Energy
noelle.metting@science.doe.gov

Lisa Miller
Brookhaven National Laboratory
lmiller@bnl.gov

Marissa Mills
Oak Ridge National Laboratory
millsmd@ornl.gov

Julie Mitchell
University of Wisconsin, Madison
mitchell@math.wisc.edu

Jennifer Morrell-Falvey
Oak Ridge National Laboratory
morrelljl1@ornl.gov

Hiep-Hoa Nguyen
TransMembrane Biosciences
hiephoa@its.caltech.edu

Sue Nokes
University of Kentucky
snokes@bae.uky.edu

Yasuhiro Oda
University of Washington
yasuhiro@u.washington.edu

Susan Old
National Institutes of Health
olds@nhlbi.nih.gov

Regina O'Neil
University of Massachusetts
rtarallo@microbio.umass.edu

Galya Orr
Pacific Northwest National Laboratory
galya.orr@pnl.gov

Mary Oster-Granite
National Institutes of Health
mo96o@nih.gov

Anna Palmisano
U.S. Department of Agriculture
apalmisano@csrees.usda.gov

Dale Pelletier
Oak Ridge National Laboratory
pelletierda@ornl.gov

Grace Peng
National Institutes of Health
penggr@mail.nih.gov

Jennifer Pett-Ridge
Lawrence Livermore National Laboratory
pettridge2@llnl.gov

Darren Platt
DOE Joint Genome Institute
dmplatt@lbl.gov

Farris Poole
University of Georgia
fpoole@bmb.uga.edu

Miguel Providenti
Environment Canada
miguel.providenti@ec.gc.ca

Chuan Qin
Scripps Research Institute
chuanqin@scripps.edu

Timothy Read
Naval Medical Research Center
readt@nmrc.navy.mil

Valerie Reed
U.S. Department of Energy
valerie.sarisky-reed@ee.doe.gov

Gemma Reguera
University of Massachusetts
greguera@microbio.umass.edu

Karin Remington
J. Craig Venter Institute
kremington@venterinstitute.org

Paul Richardson
DOE Joint Genome Institute
mcgowan4@llnl.gov

Monica Riley
Marine Biological Laboratory
mriley@mbl.edu

Jorge Rodrigues
Michigan State University
rodrig76@msu.edu

Margaret Romine
Pacific Northwest National Laboratory
Imalyn.Knight

Nagiza Samatova
Oak Ridge National Laboratory
samatovan@ornl.gov

Venkata Saripalli
Pacific Northwest National Laboratory
Ratna.Saripalli@pnl.gov

Herbert Sauro
Keck Graduate Institute
hsauro@kgi.edu

Marianne Schiffer
Argonne National Laboratory
mschiffer@anl.gov

Denise Schmoyer
Oak Ridge National Laboratory
schmoyerdd@ornl.gov

Jorg Schwender
Brookhaven National Laboratory
schwender@bnl.gov

Bari Scott
SoundVision Productions
bari@svproductions.org

Mark Segal
Environmental Protection Agency
segal.mark@epa.gov

Margrethe Serres
Marine Biological Laboratory
mserres@mbl.edu

Lucy Shapiro
Stanford University
shapiro@stanford.edu

Douglas Sheeley
National Institutes of Health
sheeleyd@mail.nih.gov

Joseph Shiloach
National Institutes of Health
Yossi@nih.gov

Michael Shuler
Cornell University
mls50@cornell.edu

Amy Shutkin
Lawrence Berkeley National Laboratory
ashutkin@lbl.gov

Steven Singer
Lawrence Livermore National Laboratory
singer2@llnl.gov

Gary Siuzdak
Scripps Research Institute
siuzdak@scripps.edu

Harold Smith
CARB/UMBI
smithh@umbi.umd.edu

Richard Smith
Pacific Northwest National Laboratory
rds@pnl.gov

Hamilton Smith
J. Craig Venter Institute
hsmith@venterinstitute.org

Jiuzhou Song
University of Maryland
songj88@umd.edu

Thomas Squier
Pacific Northwest National Laboratory
thomas.squier@pnl.gov

Ranjan Srivastava
University of Connecticut
srivasta@engr.uconn.edu

David Stahl
University of Washington
dastahl@u.washington.edu

Marvin Stodolsky
U.S. Department of Energy
marvin.stodolsky@science.doe.gov

Sergey Stolyar
University of Washington
sstolyar@u.washington.edu

Gary Stormo
Washington University Medical School
stormo@genetics.wustl.edu

Micheline Strand
U.S. Army Research Office
micheline.strand@us.army.mil

F. William Studier
Brookhaven National Laboratory
studier@bnl.gov

John Sutherland
Brookhaven National Laboratory
jcs@bnl.gov

John Tainer
Scripps Research Institute
jat@scripps.edu

Michael Thelen
Microbial Systems Division
mthelen@llnl.gov

David Thomas
J. Craig Venter Institute
dthomas@venterinstitute.org

David Thomassen
U.S. Department of Energy
david.thomassen@science.doe.gov

James Tiedje
Michigan State University
tiedjej@msu.edu

Sunia Trauger
Scripps Research Institute
strauger@scripps.edu

Emily Turner
University of Washington
emilyt@u.washington.edu

Edward Uberbacher
Oak Ridge National Laboratory
ube@ornl.gov

Clifford Unkefer
Los Alamos National Laboratory
cju@lanl.gov

Pat Unkefer
Los Alamos National Laboratory
punkefer@lanl

Ravishankar Vallabhajosyula
Keck Graduate Institute
rrao@kgi.edu

Daniel van der Lelie
Brookhaven National Laboratory
vdlelied@bnl.gov

Sanjay Vashee
J. Craig Venter Institute
svashee@venterinstitute.org

Nathan VerBerkmoes
Oak Ridge National Laboratory
nve@ornl.gov

Wim Vermaas
Arizona State University
wim@asu.edu

Patrick Viollier
Case Western Reserve University
patrick.viollier@case.edu

Peter Walian
Lawrence Berkeley National Laboratory
PJWalian@lbl.gov

Judy Wall
University of Missouri, Columbia
wallj@missouri.edu

Masakatsu Watanabe
University of California, Merced
mwatanabe@ucmerced.edu

Steven Watt
OpGen, Inc.
swatt@opgen.com

Sharlene Weatherwax
U.S. Department of Energy
sharlene.weatherwax@science.doe.gov

Peter Weber
Lawrence Livermore National Laboratory
weber21@llnl.gov

H. Steven Wiley
Pacific Northwest National Laboratory
steven.wiley@pnl.gov

Ying Xu
University of Georgia
xyn@bmb.uga.edu

Yunfeng Yang
Oak Ridge National Laboratory
yangy@ornl.gov

Fan Yang
Michigan State University
yangfan1@msu.edu

Steven Yannone
Lawrence Berkeley National Laboratory
SMYannone@lbl.gov

Jane Ye
National Heart, Lung, and Blood Institute
yej@nhlbi.nih.gov

Malin Young
Sandia National Laboratories
mmyoung@sandia.gov

Lei Young
Venter Institute
lyoung@ozex.biz

Kun Zhang
Harvard Medical School
kzhang@genetics.med.harvard.edu

Jizhong Zhou
University of Oklahoma/Institute of Environmental
Genomics
jzhou@ou.edu

Appendix 2: Web Sites

Program Web Sites

- Genomics:GTL Web site: http://doegenomestolife.org
- This book: http://doegenomestolife.org/pubs/2006abstracts/
- DOE Microbial Genome Program: http://microbialgenome.org

Project and Related Web Sites

- 3D Time-Course Visualization: http://public.kgi.edu/~fbergman/Simulate3D.htm
- Argonne National Laboratory, Biosciences Division, Combinatorial Biology: http://www.bio.anl.gov/combinatorialbiology/GTL.htm
- BACTER Institute: http://www.bacter.wisc.edu
- BioCyc Database Collection: http://biocyc.org
    - Pathway Tools Omics Viewer: http://biocyc.org/ov-expr.shtml
    - BioCyc Guided Tour: http://biocyc.org/samples.shtml
- BioWarehouse System: http://bioinformatics.ai.sri.com/biowarehouse/
- Carbon Sequestration in *Synechococcus sp.*: From Molecular Machines to Hierarchical Modeling: http://www.genomes2life.org
- Center for Molecular and Cellular Systems: http://www.ornl.gov/sci/GenomestoLife/
    - Endogenous Pulldown Approach: http://maple.lsd.ornl.gov/cgi-bin/gtl_demo/public_target_status.cgi
    - Exogenous Pulldown Approach: http://maple.lsd.ornl.gov/cgi-bin/gtl_demo/public_ex_target_status.cgi
    - Mass Spectrometry Analysis Results: http://maple.lsd.ornl.gov/gtl_demo/index.html
- Computational Systems Biology at Keck Graduate Institute; Systems Biology Workbench: http://www.sys-bio.org
- CyanoSeed (Cyanobacteria from SEED): http://theseed.uchicago.edu/FIG/organisms.cgi?show=cyano
- CyanoBase *Synechocystis*: http://www.kazusa.or.jp/cyanobase/Synechocystis/index.html
- Genome Channel Organism Selection, Microbial Genomes: http://genome.ornl.gov/microbial/
- J. Craig Venter Institute: http://www.venterinstitute.org
- Joint Genome Institute: http://www.jgi.doe.gov
- METLIN Metabolite Database: http://metlin.scripps.edu

Web Sites

- MicrobesOnline: http://microbesonline.org
- Microbial Ecology, Proteogenomics & Computational Optima. DOE Genomes to Life Center at Harvard/MIT/BWH/MGH: http://arep.med.harvard.edu/DOEGTL/
- Microbial Encyclopedia: http://modpod.csm.ornl.gov/gtl/
    - *Shewanella* Knowledgebase: http://modpod.csm.ornl.gov/shew/
- *Ralstonia metallidurans* Draft Sequence http://genome.jgi-psf.org/draft_microbes/ralme/ralme.home.html
- Ribosomal Database Project-II: http://rdp.cme.msu.edu
- *Shewanella* Federation: http://shewanella.org
- *Shewanella* Federation Integrated Knowledge Resource http://MSProRata.org (not yet online)
- Sorcerer II Expedition: http://www.sorcerer2expedition.org
- Taxonomy Browser: http://taxoweb.mmg.msu.edu
- University of Massachusetts, Amherst: http://geobacter.org/gtl/
- Virtual Institute for Microbial Stress and Survival: http://vimss.lbl.gov
    - VIMSS Repository for Experimental Data: http://vimss.lbl.gov/~jsjacobsen/cgi-bin/GTL/VIMSS/datarepository.cgi
    - VIMSS Biofiles database: http://vimss.lbl.gov/perl/biofiles/ [password required]
    - Knockout Mutants of Shewanella MR1: https://vimss.lbl.gov/~jsjacobsen/cgi-bin/Test/HazenLab/Omnilog/home.cgi [password required]

Author Index

Indexed by page number.

Indexed by page number.

Indexed by page number.

Indexed by page number.

Indexed by page number.

Indexed by page number.

Indexed by page number.

Institutional Index

Indexed by page number.

Indexed by page number.