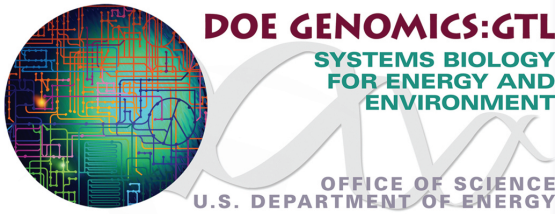
DOE/SC-0103



Joint Meeting

Genomics:GTL Awardee Workshop VI

and

Metabolic Engineering Working Group

Interagency Conference on Metabolic Engineering 2008

Bethesda, Maryland
February 10–13, 2008

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://genomicsgtl.energy.gov

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

Welcome to the 2008 Joint Genomics:GTL Awardee Workshop and the eighth Metabolic Engineering Interagency Working Group Conference. The Genomics:GTL program supports fundamental research on microbes and plants with an emphasis on understanding systems biology across multiple scales of organization. Molecular interactions among proteins, regulatory networks and metabolic pathways of individual organisms, and multicellular interactions in complex microbial communities are explored using advanced molecular and computational biology approaches enabled by genome sequencing. Research supported by the Genomics:GTL program addresses critical DOE missions in bioenergy, bioremediation of environmental contaminants, and biogeochemical cycling and biosequestration of carbon.

The past year has seen many exciting new developments for the Genomics:GTL program. In June of 2007, Secretary of Energy Samuel Bodman announced the funding of three multidisciplinary Bioenergy Research Centers aligned with the GTL program. Each center represents a multidisciplinary, collaborative effort between DOE national laboratories, universities, and private companies aimed at improving digestibility of lignocellulosic biomass from bioenergy feedstocks, discovery and bioengineering of new microbes and enzymatic systems capable of breaking down cellulose, and conversion of cellulose-derived sugars to ethanol or other biofuels. Genomics:GTL continues to support groundbreaking research by individual investigators and interdisciplinary research teams, and new efforts in technology development for imaging lignocellulose degradation, validation of genome sequence annotations, characterizing complex microbial communities, and quantitative biochemistry and metabolic engineering for biological hydrogen production were funded in 2007. The Genomics:GTL program in Ethical, Legal, and Societal Issues (ELSI) also continued to expand in 2007, incorporating new projects that address societal and sustainability issues associated with bioenergy development. The diverse array of approaches represented by these new projects and the existing Genomics:GTL community form a robust and highly complementary research program that engages some of our most pressing national priorities.

For the third year, this meeting brings together researchers supported by the Genomics:GTL program and the Interagency Metabolic Engineering Working Group. The goal of the Metabolic Engineering Working Group is the targeted and purposeful alteration of metabolic pathways found in an organism in order to better understand and use cellular pathways for chemical transformation, energy transduction, and supramolecular assembly. In addition to overlapping technological approaches, these two programs share an underlying conceptual goal of advancing understanding of organisms at the systems level.

We look forward to an exciting and productive meeting and encourage you to exchange ideas and share your expertise with other researchers. We thank you for lending your knowledge, creativity, and vision to Genomics:GTL and the Metabolic Engineering Working Group and wish you continued success in the coming year.

Jerry Elwood
Acting Associate Director of Science for
Biological and Environmental Research
Office of Science
U.S. Department of Energy

Michael Strayer
Associate Director of Science for
Advanced Scientific Computing Research
Office of Science
U.S. Department of Energy

Fred Heineken
Chair of the Interagency Metabolic
Engineering Working Group
National Science Foundation

See page 169 for additional abstracts and agenda. Addendum abstracts not included in index.

Contents

* Presenting author

Biofuels > Metabolic Engineering for Biofuels Production

BioHydrogen > Quantitative Microbial Biochemistry and Metabolic Engineering for Biological Hydrogen Production

Systems Environmental Microbiology

* Presenting author

* Presenting author

* Presenting author

Molecular Interactions and Protein Complexes

Validation of Genome Sequence Annotation

Addendum of late abstracts and agenda available starting page 169. Late abstracts not indexed.

* Presenting author

Introduction to Workshop Abstracts

Genomics:GTL Goal and Objectives

Ultimate Scientific Goal
Achieve a predictive, systems-level understanding of plants, microbes, and biological communities, via integration of fundamental science and technology development, to enable biological solutions to DOE mission challenges, including energy, environment, and climate.

Objective 1: Determine the genomic properties, molecular and regulatory mechanisms, and resulting functional potential of microbes, plants, and biological communities central to DOE missions.

Objective 2: Develop the experimental capabilities and enabling technologies needed to achieve a genome-based, dynamic systems-level understanding of organism and community functions.

Objective 3: Develop the knowledgebase, computational infrastructure, and modeling capabilities to advance the understanding, prediction, and manipulation of complex biological systems.

Abstract Organization

Abstracts associated with the Metabolic Engineering Working Group (MEWG) are identified as such and are intermixed with GTL abstracts in relevant categories. The Genomics:GTL and MEWG program abstracts and posters are organized according to the following research areas important to achieving the ultimate GTL scientific goal and objectives.

Systems Biology for DOE Energy and Environmental Missions

Bioenergy

- Biofuels: Bioenergy Research Centers
- Biofuels: Analytical and Imaging Technologies for Studying Lignocellulosic Material Degradation
- Biofuels: Metabolic Engineering for Biofuels Production
- BioHydrogen: Quantitative Microbial Biochemistry and Metabolic Engineering for Biological Hydrogen Production

Systems Environmental Microbiology

Systems Biology Research Strategy and Technology Development

Genomic and Proteomic Strategies

Molecular Interactions and Protein Complexes

Validation of Genome Sequence Annotation

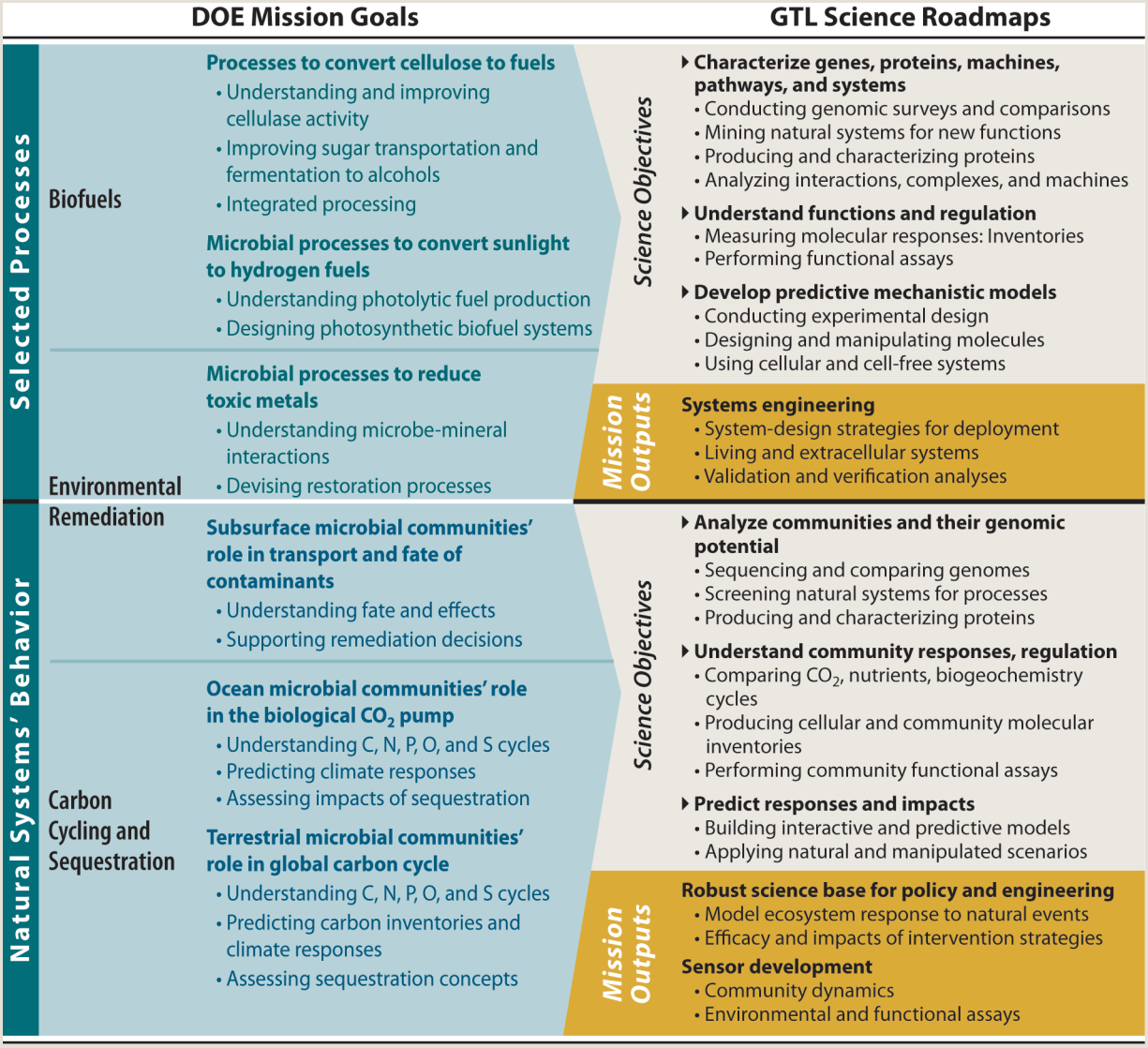
Computing Resources and Databases

Communication

Ethical, Legal, and Societal Issues

The following table is a summation of how GTL science and DOE missions align (*DOE Genomics:GTL Roadmap: Systems Biology for Energy and Environment*; October 2005, p. 40). (GenomicsGTL.energy.gov)

Summary Table. GTL Science Roadmap for DOE Missions



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.

Systems Biology for DOE Energy and Environmental Missions

Bioenergy

Biofuels > Bioenergy Research Centers

1 GTL

Great Lakes Bioenergy Research Center

Timothy Donohue,[1]* Kenneth Keegstra,[2] Richard Amasino,[1] Bruce Dale,[2] Robert Landick,[1] John Ohlrogge,[1] George Phillips,[2] Phillip Robertson,[2] and Michael Sussman[1]

[1]University of Wisconsin, Madison, Wisconsin and [2]Michigan State University, East Lansing, Michigan

Project Goals: Remove the bottlenecks in converting biomass into biofuels.

The Great Lakes Bioenergy Research Center (GLBRC) is led by the University of Wisconsin-Madison, in close partnership with Michigan State University (www.greatlakesbioenergy.org).

Additional scientific partners include the Pacific Northwest National Laboratory, Oak Ridge National Laboratory, University of Florida, Illinois State University, Iowa State University, and Lucigen Corporation. Located in the world's most productive agricultural region, the GLBRC is taking scientifically diverse approaches to converting sunlight and diverse plant feedstocks (agricultural residues, wood chips, grasses) into fuels. GLBRC programs are organized in the following areas:

Improving Plant Biomass: In addition to studying how genes affect cell-wall assembly, GLBRC will be breeding plants to produce hemicellulose, starches and oils that are more easily processed into fuels. GLBRC researchers aim to increase the energy density of grasses and other non-traditional oil crops by understanding and manipulating the metabolic and genetic circuits that control accumulation of oils and other easily digestible polymers in plant tissues.

Improving Biomass Processing: New and improved catalysts created by the GLBRC protein-production pipeline will be used with a range of plant materials and pretreatment conditions to identify the best mix of enzymes, chemicals, and physical processing for enhanced digestibility and increased fuel production from specific biomass sources. To decrease the costs of producing and using enzymes to breakdown plant cellulose, researchers are working to express biomass-degrading catalysts in the stems and leaves of plants—essentially designing crops that "self-destruct" on cue in the biofuel-production facility.

Improving Biomass Conversion to Energy Compounds: GLBRC biomass-conversion research is driven by the need to increase the quantity, diversity, and efficiency of energy products derived from plant biomass. Cellulosic ethanol is a major focus of GLBRC research, but the center will also improve biological and chemical methods for converting plant material into hydrogen, electricity, or other chemicals that can replace fossil fuels.

Fostering Sustainable Bioenergy Practices: For a bioeconomy to positively impact the energy grid, complex issues in agricultural, industrial, and behavioral systems must be addressed. To create a better understanding of what will ultimately influence the direction and acceptance of new bioenery technologies, GLBRC scientists will examine the environmental and socioeconomic dimensions of converting biomass to biofuel. To determine the best practices for biofuel production, GLBRC researchers will study how to minimize energy and chemical inputs for bioenergy crop production, reduce greenhouse gas emissions from the biofuel production life cycle, and predict environmental impacts of removing stalks, stems, and leaves from food crops. GLBRC scientists also will study the social and/or financial incentives needed to adopt these best practices.

Creating Technologies to Enable Advanced Bioenergy Research: Core GLBRC activities will provide cutting-edge, genome-based technologies to enable the innovative discoveries and creative solutions needed to advance bioenergy research. GLBRC researchers will deploy high-throughput, automated screens for genes and proteins in plants and microbes that affect biomass and

* Presenting author

biofuel production; integrate information from multiple research approaches; and develop predictive models for relevant enzymes, pathways, or networks that can guide the development of new plants, enzymes, or microbes that would enhance a biofuel-production pipeline.

Education and Outreach: With a history of excellence in the land-grant missions of education, training, and outreach, the GLBRC is committed to training the bioenergy leaders of tomorrow while removing today's bottlenecks in the biofuels pipeline. The GLBRC academic partners will offer new bioenergy-focused summer research programs, seminars, special courses and labs. By working with existing university programs, GLBRC scientists will develop workshops and educational modules for K–12 teachers on carbon chemistry, sustainability, and biodiversity issues related to biofuel production. GLBRC researchers will also develop general education materials and host public forums to raise awareness of and generate support for biofuels among farmers and communities.

2

GTL

The Joint BioEnergy Institute (JBEI): Biomass Conversion to Alternative Transportation Fuels

Jay Keasling[1,2]* (jdkeasling@lbl.gov), Harvey Blanch,[1,2] Wolf Frommer,[3] Blake Simmons,[4] Paul Adams,[1] and Kathe Andrews-Cramer[5]

[1]Lawrence Berkeley National Laboratory, Berkeley, California; [2]University of California, Berkeley, California; [3]Carnegie Institution for Science, Stanford, California; [4]Sandia National Laboratory, Livermore, California; and [5]Sandia National Laboratory, Albuquerque, New Mexico

Project Goals: This BioEnergy Research Center's program is directed at conversion of lignocellulosic biomass to transportation fuels.

Today, carbon-rich fossil fuels, primarily oil, coal and natural gas, provide 85% of the energy consumed in the United States. The release of greenhouse gases from these fuels has spurred research into alternative, non-fossil energy sources. Lignocellulosic biomass is renewable resource that is carbon-neutral, and can provide a raw material for alternative transportation fuels. Plant-derived biomass contains cellulose, which is difficult to convert to monomeric sugars for production of fuels. The development of cost-effective and energy-efficient processes to transform the cellulosic content of biomass into fuels is hampered by significant roadblocks, including the lack of specifically developed energy crops, the difficulty in separating biomass components, the high costs of enzymatic deconstruction of biomass, and the inhibitory effect of fuels and processing byproducts on organisms responsible for producing fuels from biomass monomers.

The Joint BioEnergy Institute (JBEI) is addressing these roadblocks in biofuels production. JBEI draws on the expertise and capabilities of three national laboratories (**Lawrence Berkeley National Laboratory (LBNL), Sandia National Laboratories (SNL)**, and **Lawrence Livermore National Laboratory (LLNL)**) two leading U.S. universities (**University of California** campuses at **Berkeley (UCB)** and **Davis (UCD)**, and the **Carnegie Institution for Science** at Stanford University to provide the scientific and technology underpinnings needed to convert the energy stored in cellulose into transportation fuels and other chemicals. Based in Emeryville, California, JBEI co-locates scientists and engineers from all the member organizations. JBEI's approach is based in three interrelated scientific divisions and a technologies division. The **Feedstocks Division** will create the knowledge required to develop improved plant energy crops to serve as the raw materials for biofuels. The **Deconstruction Division** will investigate the conversion of this lignocellulosic plant material to usable forms of sugars and aromatics. The **Fuels Synthesis Division** will create microbes that can efficiently convert sugar and aromatics into ethanol, butanol and advanced biofuels. JBEI's cross-cutting **Technologies Division** will develop and optimize a set of enabling technologies—including high-throughput, chip-based and 'omics platforms, tools for synthetic biology, multi-scale imaging facilities, and integrated data analysis. This division thus supports and integrates the scientific programs.

The objectives and approaches of JBEI's divisions will be described, together with the initial research accomplishments of each of JBEI's divisions.

JBEI web site: www.jbei.org

3

GTL

The BioEnergy Science Center: An Overview

Martin Keller* (kellerm@ornl.gov)

Oak Ridge National Laboratory, Oak Ridge, Tennessee

Project Goals: By combining engineered plant cell walls to reduce recalcitrance with new biocatalysts

* Presenting author

to improve deconstruction, BESC within five years will revolutionize the processing of biomass. These breakthroughs will be achieved with a systems biology approach and new high-throughput analytical and computational technologies to achieve (1) targeted modification of plant cell walls to reduce their recalcitrance (using *Populus* and switchgrass as high-impact bioenergy feedstocks), thereby decreasing or eliminating the need for costly chemical pretreatment; and (2) consolidated bioprocessing, which involves the use of a single microorganism or microbial consortium to overcome biomass recalcitrance through single-step conversion of biomass to biofuels.

http://www.bioenergycenter.org

The challenge of converting cellulosic biomass to sugars is the dominant obstacle to cost-effective production of biofuels in sustained quantities capable of impacting U.S. consumption of fossil transportation fuels. The BioEnergy Science Center (BESC) research program will address this challenge with an unprecedented interdisciplinary effort focused on overcoming the recalcitrance of biomass. In addition to Oak Ridge National Laboratory (ORNL), the BESC core team consists of the University of Georgia, the Georgia Institute of Technology, the University of Tennessee, the National Renewable Energy Laboratory, Dartmouth College, the Samuel Roberts Noble Foundation, and industrial partners ArborGen, Verenium, and Mascoma. Other individual PIs complete the team. The home base of BESC will be the Joint Institute for Biological Sciences building, funded by the state of Tennessee and occupied in December 2007. Located on the ORNL campus, the facility is designed specifically for interdisciplinary bioenergy research using systems biology tools. Other BESC anchor facilities include the University of Georgia's Complex Carbohydrate Research Center with extensive carbohydrate analytical and plant science expertise, the National Renewable Energy Laboratory's unique capabilities in comprehensive biomass analysis and bioprocessing, and ORNL's National Leadership Computing Facility.

By combining engineered plant cell walls to reduce recalcitrance with new biocatalysts to improve deconstruction, BESC within five years will revolutionize the processing of biomass. These breakthroughs will be achieved with a systems biology approach and new high-throughput analytical and computational technologies to achieve (1) targeted modification of plant cell walls to reduce their recalcitrance (using *Populus* and switchgrass as high-impact bioenergy feedstocks), thereby decreasing or eliminating the need for costly chemical pretreatment; and (2) consolidated bioprocessing, which involves the use of a single microorganism or microbial consortium to overcome biomass recalcitrance through single-step conversion of biomass to biofuels.

Within five years the Center will remove biomass recalcitrance as a barrier to cost-effective biofuels production by achieving a minimum two-fold reduction in the projected cost of processing for conversion of biomass to ethanol. Through this effort we will greatly enhance our understanding of cell wall structure during synthesis and conversion. The data generated will be made available through a Web portal in order to support and catalyze the bioenergy research community. The benefits of the basic research will extend beyond the five-year program by laying the foundation for developing other biomass sources and fuel products, improving productivity of switchgrass and poplar, and ensuring sustainability of lignocellulosic biofuel production.

This talk will provide an overview of the BESC start-up activities and some initial results.

4

GTL

BESC: Biomass Formation and Modification: *Populus*, A Case Study

G. Tuskan[1]* (tuskanga@ornl.gov), T. Tschaplinski,[1] U. Kalluri,[1] T. Yin,[1] L. Gunter,[1] S. Jawdy,[1] X. Yang,[1] R. Priya,[2] X. Zhang,[1] M. Cheng,[2] F. Chen,[3] S. DiFazio,[4] G. Slavov,[4] M. Hinchee,[5] D. Mohnen,[6] A. Darvill,[6] and **M. Keller**[1]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [2]University of Tennessee, Knoxville, Tennessee; [3]Samuel Roberts Nobel Foundation, Ardmore, Oklahoma; [4]West Virginia University, Morgantown, West Virginia; [5]Arborgen, Inc., Summerville, South Carolina; and [6]University of Georgia, Athens, Georgia

Project Goals: The overriding goal of the Center is to engineer wall biosynthesis pathways to generate plants with less recalcitrant walls that are deconstructed effectively by improved consolidated microbial bioprocessing, ultimately achieving optimal sugar release and conversion from the biomass with minimal or no chemical pretreatment. These efforts will be supported by the development of chemical and molecular models that predict how wall structure, recalcitrance, and biocatalyst-biomass interactions are related and by experimental and theoretical approaches to refine these models.

* Presenting author

This poster illustrates our approach to understanding Biomass formation transformation using *Populus*, which along with switchgrass are our primary feedstock targets. The goal of this area is to develop a thorough understanding of the genetics and biochemistry of plant cell wall biosynthesis so the process can be modified to reduce biomass recalcitrance. Preliminary data supporting our hypothesis that modification of cell wall structure will result in lower recalcitrance of biomass have been shown in alfalfa.

The *Populus* genome was sequenced, assembled and annotated in 2006. As a result, there have been numerous genetic and genomics tools developed for the discovery of genes and gene function in *Populus*. These tools include saturated genetic maps, whole-genome microarrays, libraries of polymorphic features within the genome, and a catalog of 45,500 predicted gene models. Moreover, there are in place extensive structured pedigrees, routine, high-throughput transformation protocols and mutated genetic lines available for study. These tools and resources will be brought to bear on the issue of the recalcitrance of plant cell walls to microbial and enzymatic deconstruction and conversion into liquid transportation fuels. Although such resources exist, we know very little about the regulation, formation and synthesis of plant cell walls.

As such, the first efforts within the *Populus* Transformation activity will be to identify all genes related to the construction of primary and secondary plant cell walls. It is estimated that as many as 5000 genes may be active during the process of cell wall formation. Genetics and genomics approaches will be used to identify 1000 genes with the largest impact on cell wall chemistry and thermochemical conversion to sugars. These 1000 genes and their effect on cell wall biosynthesis, recalcitrance and sugar release will then be tested and validated in over- and under-expression transgenic lines. An annotated and curated plant wall biosynthesis gene database will be created and the database will be used to construct cell wall biosynthesis pathways to facilitate future approaches to reduce recalcitrance of plant biomass.

The goal of the BESC *Populus* Transformation effort is to create improved plant lines that express enhanced biomass production and reduced recalcitrance, i.e. plants that are more easily deconstructed for biofuel production. Our specific goals are to: 1) identify and characterize genes involved in cell wall biosynthesis and structure and to establish which genes alter biomass recalcitrance and 2) determine how changes in cell wall composition and structure affect plant recalcitrance and to begin to develop a systems biology understanding of the genetic basis for wall structure and recalcitrance. A detailed experimental approach will be presented.

5

GTL

BESC: Biomass Deconstruction and Conversion: A Systems Biology Analysis of Biomass Ethanol from *Clostridium thermocellum*

Babu Raman,[1] Jonathan R. Mielenz,[1] Catherine K. McKeown,[1] Miguel Rodriguez Jr.,[1] Steven D. Brown,[1] Patricia K. Lankford,[1] Chongle Pan,[1] Gregory B. Hurst,[1] L. Lynd,[2] and **M. Keller**[1]* (kellerm@ornl.gov)

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee and [2]Dartmouth College, Hanover, New Hampshire

Project Goals: The overriding goal of the Center is to engineer wall biosynthesis pathways to generate plants with less recalcitrant walls that are deconstructed effectively by consolidated microbial bioprocessing, ultimately achieving optimal sugar release and conversion from the biomass with minimal or no chemical pretreatment. These efforts will be supported by the development of chemical and molecular models that predict how wall structure, recalcitrance, and biocatalyst-biomass interactions are related and by experimental and theoretical approaches to refine these models.

This poster illustrates part of our approach in Biomass deconstruction and Conversion using an example of the analysis of *Clostridia thermocellum*. The conversion goal is to develop an understanding of enzymatic and microbial biomass deconstruction, characterize and mine biodiversity, and use this knowledge to develop superior biocatalysts for consolidated bioprocessing (CBP). Four biologically mediated events occur in fuel production from biomass featuring enzymatic hydrolysis—saccharolytic enzyme production, hydrolysis, fermentation of six-carbon sugars, and fermentation of five-carbon sugars. Although these events are accomplished in several process steps in near-term designs, they could in principle be combined into a single step mediated by one microbe or microbial community. Recent data suggest a microbe-enzyme-substrate synergy results from this consolidation. A CBP approach has been described as the "ultimate low-cost configuration for cellulose hydrolysis and fermentation." Development of CBP will result in the largest projected cost reduction of all R&D-driven improvements. One of the leading candidate microorgan-

* Presenting author

isms for development of this CPB process is *Clostridium thermocellum*.

In this study, we used microarray technology to probe the genetic expression of *C. thermocellum* ATCC 27405 during cellulose and cellobiose fermentation. We also used multidimensional LC-MS/MS technology and ^{15}N-metabolic labeling strategy to quantify changes in cellulosomal proteins in response to various carbon sources (cellobiose, amorphous/crystalline cellulose (avicel) and combinations of avicel, pectin and xylan). Transcriptomic analysis involved a time-course analysis of gene expression to identify gene clusters with similar temporal patterns in expression during cellulose fermentation. Broadly, genes involved in energy production, translation, glycolysis and amino acid, nucleotide and coenzyme metabolism displayed a progressively decreasing trend in gene expression. In comparison, genes involved in cell structure and motility, chemotaxis, signal transduction, transcription and cellulosomal genes showed an increasing trend in gene expression. Proteomic analysis identified over 50 dockerin- and 6 cohesin-module containing components, including 20 *new* subunits. The list included several proteins of potential interest that specifically respond to the presence of 'non-avicel' substrates in the culture medium. Quantitative proteomic results also highlighted the importance of glycoside hydrolase (GH) family 9 enzymes in crystalline cellulose hydrolysis. Overall, the transcriptomic and proteomic results suggest a well-coordinated temporal and substrate-specific regulation of cellulosomal composition in *C. thermocellum*.

6

GTL

BESC: Characterization and Modeling: the Biomass HTP Characterization Pipeline for Assessing Improved Cell Walls and Enzymes

M. Davis[1]* (mark_davis@nrel.gov), W. Adney,[1] S. Decker,[1] M. Himmel,[1] S. Ding,[1] C. Wyman,[2] A. Ragauskas,[3] E. Uberbacher,[4] W. York,[5] B. Davison,[4] and **M. Keller**[4]

[1]National Renewable Energy Laboratory, Golden, Colorado; [2]University of California, Riverside, California; [3]Georgia Institute of Technology, Atlanta, Georgia; [4]Oak Ridge National Laboratory, Oak Ridge, Tennessee; and [5]University of Georgia, Athens, Georgia

Project Goals: The overriding goal of the Center is to engineer wall biosynthesis pathways to generate plants with less recalcitrant walls that are deconstructed effectively by consolidated microbial bioprocessing, ultimately achieving optimal sugar release and conversion from the biomass with minimal or no chemical pretreatment. These efforts will be supported by the development of chemical and molecular models that predict how wall structure, recalcitrance, and biocatalyst-biomass interactions are related and by experimental and theoretical approaches to refine these models.

BESC will pursue the following aims in three research focus areas covered in three posters. The other focus areas will develop improved plant materials and CBP methods that facilitate cost-effective conversion of biomass to fermentable sugars. The strength of BESC is the cross-cutting integration of diverse experimental, theoretical, and computational approaches. This integration and shared analysis is the mission of the Characterization and Modeling Focus area. For example, HTP physical characterization of biomass that is being screened for decreased recalcitrance will provide the basis for subsequent data mining that can reveal previously unknown correlations between recalcitrance and biomass structure. Such correlations will invariably lead to new hypotheses, which can be tested by integrating experimental and model-building approaches.

This poster will illustrate our establishment of a high-throughput (HTP) pretreatment and characterization pipeline that enables study of the structure, composition, and deconstruction of biomass to elucidate the underlying causes of recalcitrance. We will share many samples and use data management techniques to allow knowledge exchange across the center. The HTP pretreatment and characterization pipeline will screen the structure, composition, and deconstruction of biomass (approximately 5,000 samples per month) and identify the most promising samples for more detailed characterization.

We will discuss the plans to screen multiple samples from at least 10,000 *Populus* and switchgrass plants to identify variants with modified recalcitrance. Initial tests on baseline samples and some other materials will be discussed.

* Presenting author

Biofuels > Analytical and Imaging Technologies for Studying Lignocellulosic Material Degradation

7 GTL

Three-Dimensional Spatial Profiling of Lignocellulosic Materials by Coupling Light Scattering and Mass Spectrometry

P.W. Bohn[1]* (pbohn@nd.edu) and J.V. Sweedler[2] (sweedler@scs.uiuc.edu)

[1]Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, Indiana and [2]Department of Chemistry, University of Illinois, Urbana, Illinois

Project Goals: The goals of the project are to: (1) develop correlated optical and mass spectrometric imaging approaches for sub-surface imaging to address the highly scattering nature of LCMs in different states of processing; (2) adapt mass spectrometric imaging protocols to enable spatially-resolved LCM characterization; (3) create surface optical and mass spectrometric measures of lignin-hemicellulose-cellulose degradation at specific processing stages; and (4) correlate the in situ optical (Raman and SH-OCT) and mass spectrometric information to generate depth-resolved maps of chemical information as a function of spatial position and processing time.

The physical and chemical characteristics of lignocellulosic materials (LCMs) pose daunting challenges for imaging and molecular characterization: they are opaque and highly scattering; their chemical composition is a spatially variegated mixture of heteropolymers; and the nature of the matrix evolves in time during processing. Any approach to imaging these materials must (1) produce real-time molecular speciation information *in situ*; (2) extract sub-surface information during processing; and (3) follow the spatial and temporal characteristics of the molecular species in the matrix and correlate this complex profile with saccharification. To address these challenges we are implementing tightly integrated optical and mass spectrometric imaging approaches. Employing second harmonic optical coherence tomography (SH-OCT) and Raman microspectroscopy (RM) provides real-time *in situ* information regarding the temporal and spatial profiles of the processing species and the overall chemical degradation state of the lignin heteropolymer; while MALDI and SIMS provide spatially-resolved information on the specific molecular species produced by pre-enzymatic processing. The goals of the approach are to: (1) develop correlated optical and mass spectrometric imaging approaches for sub-surface imaging to address the highly scattering nature of LCMs in different states of processing; (2) adapt mass spectrometric imaging protocols to enable spatially-resolved LCM characterization; (3) create surface optical and mass spectrometric measures of lignin-hemicellulose-cellulose degradation at specific processing stages; and (4) correlate the *in situ* optical (Raman and SH-OCT) and mass spectrometric information to generate depth-resolved maps of chemical information as a function of spatial position and processing time.

8 GTL

Identify Molecular Structural Features of Biomass Recalcitrance Using Non-Destructive Microscopy and Spectroscopy

Shi-You Ding[1]* (shi_you_ding@nrel.gov), John Baker,[1] Mike Himmel,[1] Yu-San Liu,[1] Yonghua Luo,[1] Qi Xu,[1] John Yarbrough,[1] Yining Zeng,[1] Marcel G. Friedrich,[2] and X. Sunney Xie[2]

[1]National Renewable Energy Laboratory, Golden, Colorado and [2]Harvard University, Cambridge, Massachusetts

Project Goals: We propose to exploit recently developed techniques to image the plant ultrastructure, map the chemistry of cell-walls, and study the time-course changes during conversion processes. These techniques are novel and developed or modified specifically for biomass characterization at the nanometer resolution. They include atomic force microscopy (AFM), solid-state NMR, small angle neutron scattering (SANS), and total internal reflection fluorescence (TIRF) microscopy. We will also use fluorescence labeling approaches to map biomass surface chemistry and to track single enzyme action in vitro and in vivo. For example, one novel technique under development is to integrate AFM with TIRF-M capability to permit us to image the cell-wall substrate at the nanometer spatial resolution and to track single cell-wall/enzyme/cellulosome-substrate interaction. The application of these advances to the study of plant cell walls, and par-

* Presenting author

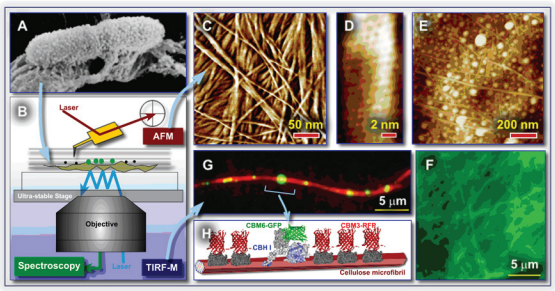
ticularly, those focused on issues relevant to biomass conversion, is currently uncharted territory.

Lignocellulosic biomass has long been recognized as a potential sustainable source of mixed sugars for fermentation to fuels and other bio-based products. However, the chemical and enzymatic conversion processes developed during the past 80 years are inefficient and expensive. The inefficiency of these processes is in part due to the lack of knowledge about the structure of biomass itself; the plant cell wall is indeed a complex nano-composite material at the molecular and nanoscales. Current processing strategies have been derived empirically, with little knowledge of the nanostructure of the feedstocks, and even less information about the molecular processes involved in biomass conversion. Substantial progress towards the cost effective conversion of biomass to fuels is contingent upon fundamental breakthroughs in our current understanding of the chemical and structural properties that have evolved in the plant cell walls which prevent its disassembly, collectively known as "biomass recalcitrance."

In nature, biomass degradation is a process of molecular interaction and reaction between plant cell wall polymers (i.e., cellulose and matrix polymers) and cellulolytic microbes and their secreted enzymes (Figure A). An integrated system (Figure B) has been set up to combine microscopic and spectroscopic modules that allow us to characterize biomass conversion processes at high spatial and chemical resolution. For example, atomic force microscopy (AFM) is used to map the surface topography of the plant cell wall and the binding of microbial cells and enzymes to the walls; total internal reflection fluorescence (TIRF) microscopy and single molecule spectroscopy is used to track the distributions and movements of labeled microbial cells and enzymes; and spectroscopy (e.g., coherent anti-stokes Raman scattering, CARS, see also Poster by *Friedrich et al.*) is used to monitor the resultant chemical changes in cell wall polymeric component during biochemical, as well as chemical, conversions of biomass.

Preliminary results have demonstrated that integrated analysis of the same cell wall samples by diverse microscopic and spectroscopic approaches is critical for characterizing the degradation processes. In the examples illustrated, correlative imaging of AFM (Figure C-E) and TIRF-M (Figure F and G) provides molecular resolution of surface structure and chemistry of cell walls.

AFM has been shown to be a powerful tool for imaging biomolecules because of its potential atomic level resolution and its ability to image surfaces under appropriately-buffered liquids (1). Precise measurement of cellulose microfibrils (Figure C) has been reported (1, 2) and individual cellulose chains (Figure D, Ding, unpublished) can be visualized. Using a flow-cell, the same cell wall sample can be thermo-chemically treated and imaged to monitor the structural changes that occur during pretreatment. Figure E shows particles precipitated on the wall surface after dilute acid pretreatment at 140°C. These particles are probably lignin-carbohydrate-complexes (LCC) generated by partial hydrolysis with acid (Ding, unpublished). The cell wall can also be specifically labeled by a fluorescently-tagged carbohydrate-binding module (CBM) and imaged by TIRF-M. Figure F shows a CBM3-GFP (green fluorescence protein) labeled cell wall, in this case, the family 3 CBM specifically recognizes the planar face of cellulose (3). This image therefore reveals cellulose distribution in the cell wall. Using similar technique, individual cellulase species can be labeled with fluorescent protein and imaged using TIRF-M at the single molecule level. Figure G shows GFP-labeled *Trichoderma* cellobiohydrolase-I (CBH I) selectively bound on areas of an individual cellulose microfibril also labeled with CBM3-RFP (red fluorescence protein) (4).



In summary, we have demonstrated non-destructive approaches to characterize the biomolecules involved in biomass conversion processes using integrated microscopy and spectroscopy. The methods are initially developed using corn stover biomass and maize plant. As a next step, we intend to employ this imaging system to characterize more energy plants, such as switchgrass and poplar wood, and their chemical and biological conversion processes to biofuels.

References

1. Ding, S.-Y. & Himmel, M. E. (2006) The Maize Primary Cell Wall Microfibril: A New Model Derived from Direct Visualization. *J. Agric. Food Chem.* 54, 597-606.
2. Himmel, M. E., Ding, S.-Y., Johnson, D. K., Adney, W. S., Nimlos, M. R., Brady, J. W., & Foust, T. D. (2007) Biomass Recalcitrance: Engineering Plants

* Presenting author

and Enzymes for Biofuels Production. *Science* 315, 804-807.

3. Xu Q., Arenkiel, P, Tucker, M. P., Ai, X., Rumbles, G., Sugiyama, J., Himmel, M. E., and Ding, S.-Y. Labeling the Planar Face of Crystalline Cellulose Using Quantum Dots Directed by Type-I Carbohydrate-Binding Modules. *Cellulose*, submitted.
4. Ding, S. Y., Xu, Q., Ali, M. K., Baker, J. O., Bayer, E. A., Barak, Y., Lamed, R., Sugiyama, J., Rumbles, G., & Himmel, M. E. (2006) Versatile derivatives of carbohydrate-binding modules for imaging of complex carbohydrates approaching the molecular level of resolution. *Biotechniques* 41, 435-443.

This work is funded in part by the U.S. Department of Energy's Genomics:GTL program under two projects, "DOE BioEnergy Science Center" and "Study of Lignocellulosic Material Degradation with CARS Microscopy".

9

GTL

Dynamic Visualization of Lignocellulose Degradation by Integration of Neutron Scattering Imaging and Computer Simulation

Barbara R. Evans[1]* (evansb@ornl.gov), Jeremy C. Smith,[2,3] Volker Urban,[1,4] Dean A. Myles,[1,4] William T. Heller,[1,4] Hugh M. O'Neill,[1] Art Ragauskas,[5] and Ida Lee[1,6]

[1]Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [2]Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, Tennessee;[3] Governor's Chair, University of Tennessee, Knoxville, Tennessee; [4]Center for Structural Molecular Biology, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [5]Institute of Paper Science and Technology, School of Chemistry and Biochemistry, Georgia Institute of Technology; and [6]Department of Electrical and Computer Engineering, University of Tennessee, Knoxville, Tennessee

Project Goals: Lignocellulosic biomass is a complex material with a molecular-level morphology and laminate structure that result in its recalcitrance to deconstruction and subsequent enzymatic hydrolysis to fermentable sugars. The objective of this research project is the development of reliable real-time in situ imaging technology integrating the various neutron and simulation techniques to obtain detailed information on the relationship of atomic and mesoscale structure to resistance to hydrolysis in lignocellulosic materials. This structural complexity will be elucidated by applying scattering techniques over a wide range of length scales (Ångströms to micrometers using wide-, small- and ultra-small angle scattering) and using several complementary approaches to contrast variation, including x-rays with sensitivity to electron density and neutrons with sensitivity to isotope composition in combination with neutron contrast variation by using appropriate blends of hydrogenated and deuterated solvents as well as by specific isotope labeling of individual biomass components. This comprehensive and multi-scale approach is necessary to address the complexity of the multi-component, multi-length-scale and multi-phase structure of lignocellulosic biomass.

Lignocellulosic biomass is composed of plant cell walls. It is a complex material composed of crystalline cellulose microfibrils laminated with hemicellulose, pectin, and lignin polymers. Understanding the physical and chemical properties of this biomass is crucial for overcoming the major technological challenge in the development of viable cellulosic bioethanol, the minimization of biomass recalcitrance to hydrolysis via the improvement of pretreatment and the design of improved feedstock plants.

To address this problem, we propose to integrate neutron science, surface force recognition imaging, and computer simulation technology to provide in situ real-time multi-scale visualization of lignocellulose deconstruction. A full arsenal of neutron scattering techniques will be employed, ranging from diffraction through small-angle scattering and dynamic neutron scattering in combination with state-of-the-art high-performance computer simulation to achieve an understanding of the physico-chemical mechanism of biomass recalcitrance in unprecedented detail. A key part is the design and employment of multipurpose neutron imaging chamber specifically designed for in situ dynamic observation under biomass pretreatment conditions. This visualization of biomass morphological degradation during heat, pressure, chemical or enzymatic treatment will provide fundamental molecular level information that will address the following basic questions:

- Visualization of regional transitions of cellulose between crystalline and amorphous.
- Investigation of lignin degradation and repolymerization in response to biomass pretreatment.
- Quantitative characterization of lignin associations with cellulose before, during and following heat, pressure, chemical or enzymatic treatment.
- Analysis of accessibility of bio-mass components to solvent molecules, chemical agents or enzymes

* Presenting author

- Visualization of structural changes of cellulose and lignocellulose during disruption and degradation by cellulases and cellulosomes.

We will present a summary of current progress on the research tasks that we have initiated to achieve these objectives.

This work is funded by the U. S. Department of Energy, Office of Biological and Environmental Research Genomics:GTL Program under FWP ERKP704, "Dynamic Visualization of Lignocellulose Degradation by Integration of Neutron Scattering Imaging and Computer Simulation."

10

GTL

A New Solution-State NMR Approach to Elucidate Fungal and Enzyme/Mediator Delignification Pathways

Kenneth E. Hammel[1]* (kehammel@wisc.edu), John Ralph,[2] Daniel J. Yelle,[1] Fachuang Lu,[2] Semarjit Shary,[1] and Sally Ralph[1]

[1]USDA Forest Products Laboratory, Madison, Wisconsin and [2]U.S. Dairy Forage Research Center, Madison, Wisconsin

Project Goals: Our goal is to use new NMR approaches to characterize the changes fungi cause in lignin when they biodegrade lignocellulose. This work will likely identify reactions that are pivotal in the removal of lignin.

The removal of lignin is a key step in lignocellulose biodegradation by fungi, and the mechanisms responsible may find practical applications in lignocellulose processing. However, the characterization of the chemical changes in lignin that fungi introduce to promote its biodegradation remains a challenge. We are using new NMR and isotope enrichment approaches to address this problem. One improvement is that lignocellulose samples undergoing biodegradation can now be completely solubilized and analyzed by two-dimensional ^{1}H-^{13}C correlation (HSQC) solution-state NMR spectroscopy. For example, by using this method we have now shown that intermonomer lignin sidechains were markedly depleted in spruce wood undergoing decay by the brown rot basidiomycete *Gloeophyllum trabeum*. Additional work is needed to characterize the products, but it is already clear, contrary to the general view, that the aromatic polymer remaining after extensive brown rot is no longer recognizable as lignin. A second improvement is that the ligninolytic reactions of some poorly understood fungi can now be characterized by ^{13}C NMR spectroscopy if the lignin in the growth substrate is enriched with ^{13}C. To test this approach, we spiked aspen wood with an α-^{13}C-labeled synthetic lignin, inoculated the wood with the soft rot ascomycete *Daldinia concentrica* to decay it, extracted the residual lignin afterwards, and analyzed this sample by solution-state ^{13}C NMR spectroscopy. The results showed that *D. concentrica* cleaved the lignin between C_α and C_β to give lignin fragments terminated by benzaldehyde and benzoic acid moieties.

11

GTL

Development of Modular Platforms for in Vivo Mapping of Local Metabolite Concentrations Important to Cell Wall Degradation by Microorganisms

Christopher G. Hunt[1]* (cghunt@fs.fed.us) and Kenneth Hammel[1,2] (kehammel@wisc.edu)

[1]USDA Forest Products Laboratory, Madison, Wisconsin and [2]Department of Bacteriology, University of Wisconsin, Madison, Wisconsin

Project Goals: Develop sensors which can be used to produce micron scale maps of metabolite activities in regions where fungi are actively degrading wood.

Background

Lignin is a major barrier to the conversion of plant cell walls into fuels or other valuable products. A complex polymer of randomly coupled phenylpropane units, lignin resists chemical degradation. It coats the cellulose and hemicelluloses of vascular plants, thus preventing enzymes from hydrolyzing these potentially valuable polysaccharides. Lignin accounts for about 25% of woody biomass and about 15% of the total biomass of agricultural feedstocks such as switchgrass. The presence of lignin is a significant hindrance to effective biomass utilization, as evidenced by the fact that fermentability of plant material is inversely related to lignin content. Cost-effective lignin removal is a significant barrier to effective fermentation of biomass.

One approach to solving the problem of lignin removal is to apply the mechanisms developed by organisms that have evolved to metabolize lignocellulose. So far, the only organisms clearly shown to biodegrade lignified tissue

* Presenting author

are certain filamentous fungi. One reason that vascular plant cell walls are so difficult to digest is that their lignin prevents even the smallest enzymes from penetrating. Filamentous fungi solve this problem by making lignocellulose more hydrophilic and porous, with the result that hydrolytic enzymes can eventually penetrate the substrate.

To make lignocellulosics permeable to enzymes, filamentous fungi use a variety of small, diffusible reactive oxygen species (ROS) such as hydroxyl radicals, peroxyl radicals, and possibly phenoxy radicals. These radicals diffuse into the cell walls and initiate biodegradative radical reactions. When lignin is the target, radical attack results in various extents of oxidation and depolymerization.

These small diffusible oxidative species are important tools used by filamentous fungi to make the cell wall accessible to enzymes. Despite this, we have a poor knowledge of how these oxidants are spatially distributed in biodegrading lignocellulose relative to the fungal hypha that produce them. The goal of this project is to remedy this deficit through fluorescence microscopy of newly designed sensors that will serve as in situ reporters of biodegradative radical production. While developing these sensors, we will test the specificity of reaction of a variety of promising fluorophores to increase the arsenal of ROS specific probes available for work at low pH. We will also demonstrate how binding these fluorescent probes to beads improves fluorescent imaging by preventing dye diffusion, limiting interferences, and allowing the use of almost all dyes (lipo- or hydrophilic, cell permeant or not) to be used in extracellular environments. Finally, we will use these sensors to produce oxidative maps that will help us to understand how fungi generate ROS and how they use these ROS to make cell walls more accessible to enzymes.

Method

We are attaching fluorescent dyes to silica beads. Our first bead has BODIPY 581/591® on a 3µm porous HPLC bead. This dye's emission changes irreversibly from red to green upon oxidation by ROS. The ratio of red to green emission provides a quantitative measure of the cumulative oxidation at that point in space. Dyes with reactivity to specific ROS, pH, or other metabolites of interest are envisioned.

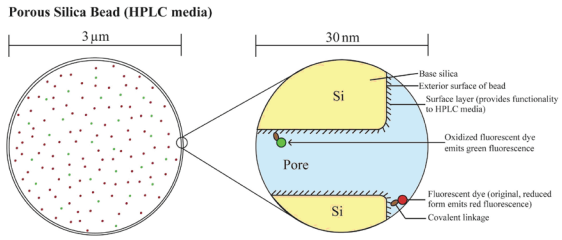


Figure. Fluorescent dye attached to silica bead

There are many advantages gained by fixing the dye to bead. We design the bead to emit two fluorescent signals, so that the ratio of the two signal intensities provides quantitative information. Immobilized dyes are prevented from moving after reaction, so partitioning is impossible, they cannot be ingested, and the fluorescence from the dye is clearly distinguishable from background.

Beads are placed on wood samples and imaged with a confocal microscope during fungal colonization. Images can be analyzed to provide the analyte concentration maps as well as an overlay of the location of fungal hypha.

12

GTL

Dynamic Molecular Imaging of Lignocellulose Processing in Single Cells

Amy Hiddessen, Alex Malkin, and **Michael Thelen*** (mthelen@llnl.gov)

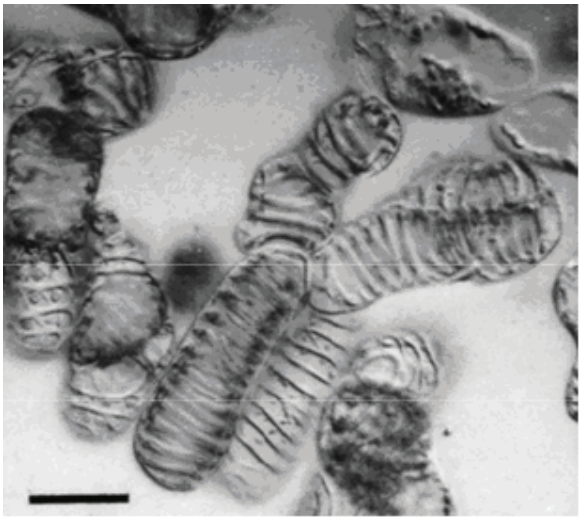
Chemistry Directorate, Lawrence Livermore National Laboratory, Livermore, California

Of central importance to our nation's energy resources is the pursuit of alternative fuels directly from plant material rich in lignin and cellulose. The abundant and intractable nature of these biopolymers limits their practical use to produce biofuel precursors, and so presents an extraordinary scientific challenge. A comprehensive understanding of the natural or engineered breakdown of plant cell wall materials must be addressed on several levels, including the examination of detailed ultrastructural changes that occur in real time. To accelerate research on the cellular and molecular details of the cell wall deconstruction process, we are developing sophisticated analytical tools specifically to visualize changes in surfaces, polysaccharides and proteins. Molecular surface characterization can be directly linked with high

* Presenting author

resolution, three-dimensional images of cellular structure by combining atomic force microscopy (AFM) with laser scanning confocal microscopy (LSCM) in a single instrument.

Highly oriented assemblies of lignocellulose are found in the water-transporting xylem tissue. To observe these thick secondary wall depositions in single cells, we are using the *Zinnia elegans* culture system in which leaf mesophyll cells are induced to differentiate in synchrony to tracheary elements (single xylem cells; see Figure). In these cultures, a high proportion of cells undergo programmed autolysis which is accompanied by a high rate of secondary wall polysaccharide synthesis and lignification. When isolated at the same stage of wall formation, cells are treated with enzyme preparations or wood-degrading microbes to stimulate specific decomposition of wall materials, and these are examined under physiological conditions using AFM for topological, near-molecular scale imaging. Corresponding processing of polymers, protein composition, and protein co-localization will be detected and visualized using gold-conjugated antibodies by AFM-based immunolabeling. Specific dyes and antibodies also will be used as cellular probes in LSCM, where live cell samples will be observed in real time in a precisely controlled environment. Methods will be developed to couple AFM and LSCM imaging techniques for monitoring natural and induced changes in wall ultrastructure, and rendered as 3D reconstructions.



Cultured cells of *Zinnia elegans*. Helically oriented secondary cell wall thickenings are characteristic of xylem cells in culture as in plant vasculature. bar = 25 µm.

13

GTL

Study of Lignocellulosic Material Degradation with CARS Microscopy

Marcel G. Friedrich[1]* (friedric@fas.harvard.edu), Conor L. Evans,[1] Gary Holtom,[1] John Yarbrough,[2] **Shi-You Ding**[2] (shi_you_ding@nrel.gov), and X. Sunney Xie[1] (Xie@chemistry.harvard.edu)

[1]Harvard University, Cambridge, Massachusetts and [2]National Renewable Energy Laboratory, Golden, Colorado

Project Goals: Finding alternative sources of energy required to maintain today's living standards is a compelling challenge. A promising substitute for fossil fuel is bioethanol produced from biomass conversion. The key step of this process is the degradation of plant cell wall polysaccharides into fermentable sugars. Lignin is the hydrophobic polymeric component of the plant cell wall. It has been proposed that lignin may play a role in preventing plant cell walls from chemical and biological hydrolysis during biomass conversion processes. The improvement of conversion efficiency is key to biofuel production and requires imaging techniques with contrast based on their chemical composition for a real-time monitoring. This project is aimed on applying coherent anti-Stokes Raman Scattering (CARS) microscopy to visualize the chemical/structural changes of the cell wall polymers during the degradation process.

Finding alternative sources of energy required to maintain today's living standards is a compelling challenge. A promising substitute for fossil fuel is bioethanol produced from biomass conversion. The key step of this process is the degradation of plant cell wall polysaccharides into fermentable sugars. Lignin is the hydrophobic polymeric component of the plant cell wall. It has been proposed that lignin may play a role in preventing plant cell walls from chemical and biological hydrolysis during biomass conversion processes. The improvement of conversion efficiency is key to biofuel production and requires imaging techniques with contrast based on their chemical composition for a real-time monitoring. This project is aimed at visualizing the chemical/structural changes of the cell wall polymers during the degradation process.

Confocal and multiphoton fluorescence microscopy have become powerful techniques for three-dimensional imaging of biological systems. However, for biochemical species or cellular components that either do not fluoresce or cannot tolerate labeling, other contrast mechanisms with molecular specificity are needed. Vibrational microscopy

* Presenting author

based on infrared absorption and spontaneous Raman scattering have been used for chemically selective imaging. However, IR microscopy is limited to low spatial resolution because of the long wavelength of light used. Furthermore, the absorption of water in the infrared region makes it difficult to image biological materials. Raman microscopy can overcome these limitations and has been successfully used to characterize plant cell walls. By tuning into different vibrational frequencies, lignin and carbohydrates could be imaged selectively. However, the intrinsically weak Raman signal can require high laser powers and long integration times, often hours. This poor time resolution prevents any kind of real time monitoring.

Coherent anti-Stokes Raman scattering (CARS) microscopy has matured as a powerful nonlinear vibrational imaging technique that overcomes these limitations of conventional Raman microscopy [3,4]. CARS microscopy provides a contrast mechanism based on molecular vibrations, which are intrinsic to the samples. It does not require natural or artificial fluorescent probes. CARS microscopy is orders of magnitude more sensitive than spontaneous Raman microscopy. Therefore, CARS microscopy permits fast vibrational imaging at moderate average excitation powers (i.e. up to ~10 mW) tolerable by most biological samples. Furthermore, CARS microscopy has a three-dimensional sectioning capability, useful for imaging thick tissues or cell structures. This is because the nonlinear CARS signal is only generated at the focus where the laser intensities are the highest. Finally, the anti-Stokes signal is blue-shifted from the two excitation frequencies, and can thus be easily detected in the presence of one-photon fluorescence background. We demonstrate that CARS can visualize the chemical and structural changes of the cell walls during this conversion process.

The Raman spectrum of a cell wall region in corn stover is shown in Fig. A (inset). The prominent band at 1600 cm^{-1} is the aryl symmetric ring stretching vibration which serves as a sensitive marker for the presence of lignin. The two bands at around 1090 and 1110 cm^{-1} are due to C-O and C-C stretching vibrations from cellulosic polymers at the cell wall. The structures of lignin and cellulose are shown in Fig. B. Figure a shows the CARS image tuned into the lignin band at 1600 cm^{-1}. A high density of the lignin is detected in the secondary cell walls. The integration time of a CARS image (Fig. d) is with 20 sec tremendously faster than compared to the Raman picture that requires about 55 min (Fig. c), while providing a comparable contrast. The fast integration time is critical of monitoring chemical changes during biomass pretreatment that usually occurs within minutes.

Currently, we are working on improving the sensitivity of the CARS technique to image cellulose, as well as spatial resolution. These initial results demonstrate the feasibility of the proposed method.

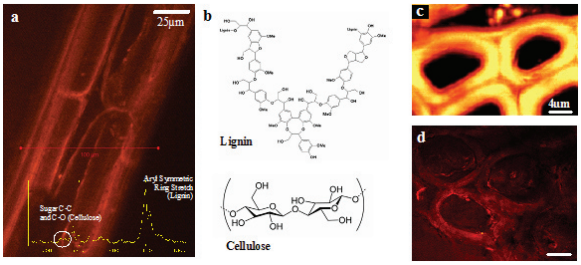


Figure. a) CARS images of cross section of corn stove cell walls by integrating over the aromatic lignin band wavelength (1,550–1,640 cm-1). b) Chemical Structures of lignin and cellulose. c) Cross section of poplar wood imaged by Raman microscopy (integration time of 55 min) and d) CARS microscopy (integration time of 20sec).

References

1. Ding, S.-Y. & Himmel, M. E. (2006) The Maize Primary Cell Wall Microfibril: A New Model Derived from Direct Visualization. *J. Agric. Food Chem.* 54, 597-606.
2. Himmel, M. E., Ding, S.-Y., Johnson, D. K., Adney, W. S., Nimlos, M. R., Brady, J. W., & Foust, T. D. (2007) Biomass Recalcitrance: Engineering Plants and Enzymes for Biofuels Production. *Science* 315, 804-807.
3. Zumbusch, Andreas; Holtom, Gary R.; Xie, X. Sunney (1999) Vibrational Microscopy Using Coherent Anti-Stokes Raman Scattering, *Phys. Rev. Lett.* 82, 4142.
4. Cheng, Ji-Xin and Xie, X. Sunney (2004) Coherent anti-Stokes Raman scattering microscopy: instrumentation, theory and applications, *J. Phys. Chem.* B 108, 827.

* Presenting author

14

GTL

Single-Molecule Studies of Cellulose Degradation by Celluosomes

Haw Yang[1,2]* (hyang@lbl.gov), Jamie Cate,[1,2] Padma Gunda,[1] L. Meadow Anderson,[1] and Hu Cang[1]

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

A cellulosome is a large extracellular supramolecular complex that is produced by anaerobic microbes to enzymatically decompose crystalline cellulosic polymers and plant cell walls. It consists of a scaffolding protein that accommodates other essential protein and enzyme components for cellulose degradation. They include carbonhydrate-binding modules for attachment to the solid cellulose substrate, various glycoside hydrolases to efficiently hydrolyze a heterogeneous substrate, and in some cases, anchoring proteins to attach cellulosomes to the bacterial cell surface. The mechanism of its function is poorly understood due to the complexity of the cellulosome itself and the natural environment in which it functions. Outstanding issues include the location and the manner in which the cellulosome is assembled, the distribution in the cellulosomal composition, and the dynamic interactions between cellulases and the insoluble substrate, to name a few.

These questions are very difficult to address quantitatively using conventional, ensemble-based methods due to the multiple layers of complexity involved. The convoluted spatio-temporal dynamics in cell and cellulosome interaction with insoluble substrates make it very hard to quantitatively study the various molecular dynamics of a functioning cellulosome. We anticipate that the single-molecule approach, due to its capability of directly monitoring the individual processes from a distribution, will prove invaluable in efforts to unravel how microscopic, molecular interactions impact macroscopic behavior in plant cell wall degradation.

In order to study the various processes involved in lignocellulosic degradation by cellulosomes, we are developing a single-molecule spectrometer with the capability to track individual fluorescent particles in three dimensions. We plan to follow the assembly and disassembly as well as the export of cellulosome complexes both *in vitro* and *in vivo*. We are developing quantitative single-molecule and single-particle assays for cellulosome activity. We are studying the cellulosome's formation and processivity for the degradation of lignocellulose. These experiments will exploit the genetic tools that we are developing to incorporate fluorescent or nanoparticle tags into the cellulosome to enable tracking. The mechanistic insights are expected to have a direct impact on the improvement and engineering of tailored biomass depolymerization systems.

Biofuels > Metabolic Engineering for Biofuels Production

15

GTL

Non-Fermentative Pathways for Synthesis of Branched-Chain Higher Alcohols as Biofuels

Shota Atsumi, Taizo Hanai, and **James C. Liao*** (liaoj@seas.ucla.edu)

Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, California

Project Goals: The goal of this project is to develop novel pathways for production of higher alcohols.

Global energy and environmental problems have stimulated increased efforts in synthesizing biofuels from renewable resources. Compared to the traditional biofuel, ethanol, higher alcohols offer advantages as gasoline substitutes because of their higher energy density and lower hygroscopicity. In addition, branched-chain alcohols have higher octane numbers compared to their straight-chain counterparts. However, these alcohols cannot be synthesized economically using native organisms. Here we present a metabolic engineering approach using *Escherichia coli* to produce higher alcohols including isobutanol, 1-butanol, 2-methyl-1-butanol, 3-methyl-1-butanol and 2-phenylethanol from a renewable carbon source, glucose. This strategy leverages the host's highly active amino acid biosynthetic pathway and diverts its 2-keto acid intermediates for alcohol synthesis. In particular, we have achieved high yield, high specificity production of isobutanol from glucose. The strategy enables the exploration of biofuels beyond those naturally accumulated to high quantities in microbial fermentation.

* Presenting author

16

SBIR

Anaerobic Expression of Pyruvate Dehydrogenase for Producing Biofuels in Fermentative Pathways

Thomas Buelter* (tbuelter@gevo.com), Renny Feldman, and Peter Meinhold

Gevo, Inc., Pasadena, California

Project Goals: Aims at determining whether a balanced butanol fermentation pathway can be expressed in a heterologous host, such as *E. coli*. The successful implementation of a metabolic pathway for the production of butanol in a heterologous host depends upon generating the necessary reducing equivalent NADH during the conversion of pyruvate to acetyl-CoA. Therefore, a heterologous host will be engineered so that it expresses pyruvate dehydrogenase, and not pyruvate formate lyase under anaerobic conditions.

Advanced biofuels, which include butanol, will deliver the performance of gasoline without the environmental impact and these biofuels will reduce our dependency on foreign oil. Butanol has a higher energy content per gallon than many first generation biofuels, it does not absorb water and can be transported through the existing oil and gas distribution infrastructure. Butanol can be used in gas-powered vehicles without modification or blending.

In order to make butanol competitive on the fuels market it has to be produced with a higher yield than currently achievable using bacteria from the genus *Clostridium*, which produces many byproducts. We are working on recombinant microorganisms that are engineered to convert biomass into butanol without byproducts. A key problem to be solved is the redox balance of the cell during butanol production from sugar. Additional reducing equivalents in the form of NADH have to be generated to produce butanol from sugar.

To accomplish this we are working on three approaches: We are using in vivo evolution and selection to select for a strain that has the ability to produce sufficient reducing power to generate butanol at high yield. The native pathway for anaerobic generation of Acetyl CoA and fermentative pathways were deleted from the strain before selection pressure was applied to only allow strains to grow that have increased NADH production. The second approach we are pursuing is the directed evolution of pyruvate dehydrogenase subunits with the aim to increase the activity of this enzyme under the conditions of butanol production. The third approach uses the engineering of the cells' regulation of pdh expression to increase its expression under butanol production conditions.

17

MEWG

Kinetic Modeling of Metabolic Pathways in *Zymomonas mobilis* to Optimize Pentose Fermentation

Dhinakar S. Kompala* (kompala@colorado.edu)

University of Colorado, Boulder, Colorado

Project Goals: To investigate the complex network of metabolic pathways in engineered bacterium *Zymomonas mobilis*. To develop a quantitative understanding of the metabolic fluxes along the newly engineered pathways for pentose fermentation.

Zymomonas mobilis has been engineered with four new enzymes to ferment xylose along with glucose. A network of pentose pathway enzymatic reactions interacting with the native glycolytic Entner Doudoroff pathway has been hypothesized. We have analyzed the complex interactions between the pentose phosphate and glycolytic pathways in this network by developing a large-scale kinetic model for all the enzymatic reactions. Based on the experimental literature on *in vitro* characterization of each of the 20 enzymatic reactions, the large-scale kinetic model is numerically simulated to predict the dynamics of all the intracellular metabolites along the network of interacting metabolic pathways. This kinetic model takes into account all the feedback and allosteric regulations on the enzymatic reaction rates and is better suited to the systems level analysis of interacting metabolic pathways compared to the standard linearized methods of metabolic flux analysis and metabolic control theory.

This nonlinear kinetic model is simulated to perform numerous *in silico* experiments by varying different enzyme concentrations and predicting their effects on all the intercellular metabolic concentrations and the ethanol production rates in continuous fermentors. Among the five enzymes whose concentrations were varied and given as input to the model, the ethanol concentration in the continuous fermentor was optimized with xylose isomerase was needed at the highest level, followed by the transaldolase. Predictions of the model, that interconnecting enzyme phosphoglucose isomerase, does not need to be overexpressed, were recently confirmed through experimental investigations. Through this kinetic modeling approach, we can develop efficient ways of maximizing the fermentation of both glucose and xylose,

* Presenting author

while minimizing the expression of the heterologous enzymes.

18

MEWG

Foundational Advances in RNA Engineering Applied to Control Biosynthesis

Christina D. Smolke* (smolke@cheme.caltech.edu)

California Institute of Technology, Pasadena, California

Project Goals: The goals of this project are to: (i) generate RNA-based molecular sensors to key metabolites along the BIA pathway; (ii) engineer *Saccharomyces cerevisiae* to synthesize early BIAs; (iii) integrate the molecular sensors into the transgenic yeast strain for the noninvasive, real-time detection of key metabolite levels along this heterologous pathway; (iv) expand the utility of the engineered molecular sensors to the dynamic regulation of enzyme levels in response to metabolite accumulation; and (v) solidify a strong base in biomolecular and metabolic engineering and encourage its advancement by training and educating scientists through cutting-edge, integrated research and educational plans.

Recent progress in developing frameworks for the construction of RNA devices is enabling rapid advances in cellular engineering applications. These devices provide scalable platforms for the construction of molecular communication and control systems for reporting on, responding to, and controlling intracellular components in living systems. Research that has demonstrated the modularity, portability, and specificity inherent in these molecules for cellular control will be highlighted and its implications for synthetic and systems biology research will be discussed. In addition, tools that translate sequence information to device function to enable the forward design and optimization of new devices will be discussed. The flexibility of the specified framework enables these molecules to be integrated as systems that perform higher-level signal processing based on molecular computation strategies. The application of these molecular devices to studying cellular systems through non-invasive *in vivo* monitoring of biomolecule levels and to regulating cellular behavior will be discussed, in particular in the control and optimization of the biosynthesis of alkaloids in *Saccharomyces cerevisiae*.

19

MEWG

Development of Tolerant and Other Complex Phenotypes for Biofuel Production

E. Terry Papoutsakis* (epaps@udel.edu)

Department of Chemical Engineering and the Delaware Biotechnology Institute, University of Delaware, Newark, Delaware

Project Goals: Develop solvent tolerant phenotypes in prokaryotes for biofuel production.

In bioprocessing, in addition to maximizing the flux for a desirable product, the robustness and prolonged productivity of the biocatalyst (the cells) under realistic bioprocessing conditions is an equally important issue. Thus, the ability of cells to withstand "stressful" bioprocessing conditions without loss of productivity is a most significant goal. Such conditions include: toxic substrates, accumulation of toxic products and byproducts, high or low pH, or high salt concentrations as encountered in most applications for the production of chemicals and biofuels as well as in bioremediation applications. The difficulty—but also the intellectual and biotechnological challenge—is that the desirable phenotypic trait is determined by several genes or a complex regulatory circuit. Complex phenotypes are also encountered when one desires to develop a *de novo* capability or pathway in a particular cell type. For example, how do cells put together the regulatory elements of a sequence of genes to make a pathway or program possible? Yes, it is an evolutionary process, but if we are to "imitate" the process, what would we do? What tools could one possibly use and strategies to facilitate the development of complex phenotypes in microbial cells? From omics-based analysis to synthesis, all selection based, or hybrid? Knowledge-based and mechanistic or not? This will be the focus of this presentation, together with some data from early efforts to demonstrate some key concepts that we explore in my laboratory.

* Presenting author

20

SBIR

Thermophilic Electricity Generating Bacteria as Catalysts for the Consolidated Bioprocessing of Cellulose

T. Shimotori,[1] C. Marshall,[2] R. Makkar,[2] M. Nelson,[1] and **H. May**[1,2]* (hmay@mfctech.net)

[1]MFC Technologies, LLC, Mount Pleasant, South Carolina and [2]Department of Microbiology and Immunology, Medical University of South Carolina, Charleston, South Carolina

Project Goals: Discover thermophilic electrode reducing bacteria that may be used with cellulolytic and ethanologenic bacteria to convert cellulose into biofuels. Apply aforementioned bacteria in modified fuel cells for the production of ethanol, electricity and hydrogen (or combination thereof) from cellulose.

The biological mediation of the conversion of lignocellulose to biofuels holds much promise, particularly when regarding the supply constraints and contribution to global warming associated with fossil fuels. Consolidated bioprocessing of cellulosic biomass leverages the catalytic activity of cellulolytic and ethanologenic bacteria to produce ethanol[1,2]. Due to their rapid and effective ability to metabolize cellulose, the bacteria utilized in this process are thermophilic, e.g. *Clostridium thermocellum* and *Thermoanaerobacter thermosaccharolyticum*. However, the formation of acetate and other by-products during the processing and fermentation of plant biomass is common, which may inhibit the overall fermentation. Therefore, it would be advantageous to use bacteria capable of consuming acetate in combination with the cellulolytic bacteria. One clever way to consume acetate under anaerobic conditions is to use electricity generating bacteria and microbial fuel cells (MFCs), where acetate is converted to carbon dioxide and electricity is generated. A thorough description of MFCs is given in a new book authored by Bruce Logan[3]. We set out to discover thermophilic electricity generating bacteria that would be compatible with the thermophilic ethanologenic bacteria with the idea that ethanol and electricity could then be generated from cellulose.

To find thermophilic electricity generating bacteria we enriched for acetate-consumers in MFCs beginning with sediment fuel cells from marine and freshwater sources incubated at 60°C. This population was further enriched in sediment-free single chamber fuel cells equipped with air cathodes. Following several exchanges of the media and several transfers to additional fuel cells with only acetate as an energy source, the enriched community from Charleston Harbor was scraped from the surface of an anode and the 16S rRNA genes of the community were cloned and sequenced. A mixed community that included bacteria most closely related to *Deferribacter* spp. but dominated by Gram positive *Firmicutes*, particularly of *Thermincola* spp., was discovered. The *Deferribacter* spp. have been isolated from deep wells and deep-sea hydrothermal vents and some species are known to reduce metals external to the cells[4-6]. Two *Thermincola* spp. have been isolated from volcanic hotsprings[7,8]. Interestingly, our community came from a mesobiotic estuarine harbor.

Spores from the *Thermincola* are resistant to autoclaving, and our enriched community would still grow and generate electricity in a fuel cell supplied with acetate even after the inoculum was autoclaved for 30 minutes. The 16S rRNA genes of the autoclaved community are now under analysis. In addition, *T. ferriacetica*, which is capable of reducing insoluble iron oxides while consuming acetate, is in pure culture and we are now testing it for the ability to generate electricity in a fuel cell. Both the enriched community and *T. ferriacetica* are also being tested in combination with cellulolytic bacteria, e.g. *C. thermocellum*, in a consolidated bioprocess to produce ethanol and electricity from cellulose.

References

1. Lynd, L.R., van Zyl, W.H., McBride, J.E. & Laser, M. Consolidated bioprocessing of cellulosic biomass: an update. *Curr. Opin. Biotechnol.* **16**, 577-583 (2005).
2. Lynd, L.R., Wiemer, P.J., van Zyl, W.H. & Pretorius, I.S. Microbial cellulose utilization: fundamentals and biotechnology. *Microbiol. Mol. Biol. Rev.* **66**, 506-577 (2002).
3. Logan, B.E. Microbial Fuel Cells. *Wiley-Interscience* (2008).
4. Miroshnichenko, M.L. et al. *Deferribacter abyssi* sp. nov., an anaerobic thermophile from deep-sea hydrothermal vents of the Mid-Atlantic Ridge. *Inter. J. Syst. Evol. Microbiol.* **53**, 1637-1641 (2003).
5. Takai, K., Kobayashi, H., Nealson, K.H. & Horikoshi, K. *Deferribacter desulfuricans* sp. nov., a novel sulfur-, nitrate- and arsenate-reducing thermophile isolated from a deep-sea hydrothermal vent. *Inter. J. Syst. Evol. Microbiol.* **53**, 839-846 (2003).
6. Greene, A.C., Patel, B.K. & Sheehy, A.J. *Deferribacter thermophilus* gen. nov., sp. nov., a novel thermophilic manganese- and iron-reducing bacterium isolated from a petroleum reservoir. *Int. J. Syst. Bact.* **47**, 505-509 (1997).

* Presenting author

7. Sokolova, T.G. et al. *Thermincola carboxydiphila* gen. nov., sp. nov., a novel anaerobic, carboxydotrophic, hydrogenogenic bacterium from a hot spring of the Lake Baikal area. *Int. J. Syst. Evol. Microbiol.* **55**, 2069-2073 (2005).

8. Zavarzina, D.G. et al. *Thermincola ferriacetica* sp. nov., a new anaerobic, thermophilic, facultatively chemolithoautotrophic bacterium capable of dissimilatory Fe(III) reduction. *Extremophiles* **11**, 1-7 (2007).

BioHydrogen > Quantitative Microbial Biochemistry and Metabolic Engineering for Biological Hydrogen Production

21

GTL

Metabolic Modeling for Maximizing Photobiological H_2 Production in Cyanobacteria

Alexander S. Beliaev[1]* (alex.beliaev@pnl.gov), Grigoriy E. Pinchuk,[1] Oleg Geydebrekht,[1] Michael H. Huesemann,[2] Jennifer L. Reed,[3] Andrei L. Osterman,[4] John R. Benemann,[5] and Jim K. Fredrickson[1]

[1]Pacific Northwest National Laboratory, Richland, Washington; [2]Pacific Northwest National Laboratory Marine Science Laboratory, Sequim, Washington; [3]University of California, San Diego, California; [4]Burnham Institute for Medical Research, La Jolla, California; and [5]Benemann Associates, Walnut Creek, California

Project Goals: In this study, we seek to improve our understanding of H_2 production by a diazotrophic unicellular cyanobacterium *Cyanothece* sp. strain ATCCC 51142 using a metabolic modeling approach for simulating the fundamental metabolism of indirect biophotolysis, as well as identifying the main metabolic and regulatory controls in this organism. From this, the potential for H_2 production by indirect biophotolysis in this organism will be assessed based on imposing new constraints to redirect the low redox potential electron transport pathways from normal metabolism towards H_2 production from accumulated carbohydrates. As a result, it will provide an in silico tool for manipulating such microorganisms to act as catalysts for solar energy conversion to H_2 and potentially allow for a development of a highly efficient H_2 production process.

Advances in microbial genome sequencing and functional genomics are greatly improving the ability to construct accurate systems-level models of microbial metabolism and to use such models for metabolic engineering. With the increasing concerns over the reliance on fossil fuels, there is a revitalized interest in using biological systems for producing renewable fuels. Genomics and metabolic engineering hold great promise for the rational design and manipulation of biological systems to make such systems efficient and economically attractive.

Although there is a relatively rich body of scientific information on the biochemistry, physiology, and genetics of photosynthetic H_2 production, a systems-level understanding of this process is lacking. In this proposal, we seek to improve our understanding of H_2 production by a diazotrophic unicellular cyanobacterium *Cyanothece* sp. strain ATCC 51142 using a metabolic modeling approach for simulating the fundamental metabolism of indirect biophotolysis as well as identifying the main metabolic and regulatory controls in this organism. The genus *Cyanothece* has several attractive properties; they are aerobic, unicellular, diazotrophic bacteria that separate in time the process of light-dependent autotrophic growth and glycogen accumulation from N_2 fixation at night. The process of N_2 fixation and concomitant cyanophycin (nitrogen storage compound) accumulation is accompanied by a decrease in total cellular glycogen content.

In this project, we intend to construct a base metabolic model from genome sequences and other sources of information and integrate new data from physiology experiments, functional genomics, and comparative sequence analysis to develop high-quality, comprehensive models suitable for predictive analysis of *Cyanothece* metabolism. From this, we will assess the potential for H_2 production by indirect biophotolysis in this organism based on imposing new regulatory constraints that would redirect the low-redox-potential electron transport pathways from normal metabolism toward H_2 production from accumulated carbohydrates. We will also extend our studies to other cyanobacterial species including *Synechocystis* PCC 6803, which will allow us to develop a more robust cyanobacterial metabolic model and conduct a comparative metabolic analysis between the two organisms. The result will be an *in silico* tool for manipulating such microorganisms to act as catalysts for solar energy conversion to H_2 and will potentially allow development of a highly efficient H_2 production process.

* Presenting author

22

GTL

Engineering Microbes for Enhanced Hydrogen Production: Parameter Estimation for Transcriptional Control of Metabolism

Alex Fichtenholtz* (alexf@bu.edu), Diogo Camacho, and **James J. Collins**

Department of Biomedical Engineering and Center for BioDynamics, Boston University, Boston, Massachusetts

Project Goals: As part of this proposed project, we plan to create a novel systems biology technology platform and resources that can be used to reverse engineer and analyze gene regulatory networks and metabolic pathways that influence and control microbial hydrogen production. We will use this novel platform to achieve the following specific objectives: (1) construct genome-scale transcriptional regulatory and metabolic models using automated model reconstruction methods, and subsequently integrate the regulatory and metabolic models for computational analysis; (2) use the integrated metabolic and regulatory models to identify relevant metabolic processes and pathways involved in hydrogen production as well as the regulatory processes that inhibit hydrogen overproduction; and (3) use the integrated metabolic and regulatory models to make computational predictions of optimal metabolic pathway modifications, which will enable enhanced biological hydrogen production. As part of this project, we plan to use our systems biology platform to study the cyanobacterium *Cyanothece*, a microbe with metabolic capabilities to produce hydrogen. We will collaborate with a team from the Pacific Northwest National Laboratory (PNNL), headed by Alexander Beliaev and involving Jim Fredrickson.

Flux distribution is a major problem in metabolic engineering, in that, while one can insert genes for a novel pathway of reactions for a metabolite into the cell, basic chemical kinetics dictates that much of the flux will flow through pre-existing reactions for this metabolite, and only some through this foreign pathway. While this approach is occasionally successful, it harbors a host of problems: (i) it is inherently a local approach, as the gene knockouts always focus on genes that surround a metabolite node of interest and (ii) such knockouts often significantly slow down the growth and nutrient uptake rates of the cell. The gene knockout strategy fails to be an acceptable solution on account of it consistently falling short of optimum efficiency and rates of production.

The methodology proposed here is to approach the problem by controlling the transcriptional network regulating the expression of metabolic genes. This network of transcription factors, which has considerable connectivity and feedback, has evolved as a response mechanism to control the cell's metabolism in a variety of conditions. It controls the distribution of flux throughout the cell on a global scale, from rate nutrient uptake, to ratio of reaction distributions at metabolite nodes, all the way down to the rate at which waste exits the cell. Several of these transcription factors, such as CRP, Fnr, and FruR, modulate the expression of dozens of metabolic genes. Optimizing the expression of these transcription factors for the best efficiency of production is a more global approach than knocking out individual metabolic genes, and thus much more likely to yield consistent results.

With our approach, the parameter that describes the strength of connection is determined and the network reconstructed through a parameter estimation of a generalized differential equation for the levels of expression for each gene. This methodology was applied to a two-reaction system composed of 19 differential equations and 60 parameters. Using synthetic time course data from both wild-type and knockout experiments, we were able to estimate an average of 75% of all parameters to within 5% precision.

Considering the success of our parameter estimation approach, an expansion of the model to the central metabolic network of *Escherichia coli* and the cyanobacterium *Cyanothece*, respectively, will be made and used in future studies to maximize hydrogen production for modified strains. Once the parameters for these gene regulatory networks are estimated, modifications to the networks can be attempted. These modifications will include a host of genetic engineering elements, such as dynamic toggle switches and riboregulators. Our reconstructed networks will allow us to quantitatively predict the effect of any modifications to the network topology, and thus to predict flux distribution throughout the central metabolic network with a high degree of accuracy.

Acknowledgements: This work was supported by the DOE under Award Number DE-FG02-07ER64483.

* Presenting author

23

GTL

High-Throughput Screening Assay for Biological Hydrogen Production

Roger L. Ely[1,2*] (ely@engr.orst.edu), Paul S. Schrader,[2,3] and Elizabeth H. Burrows[1]

[1]Department of Biological and Ecological Engineering, Oregon State University, Corvallis, Oregon; [2]Department of Chemical Engineering, Yale University, New Haven, Connecticut; and [3]Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon

Project Goals (Abstracts 23-27): The objectives of the research are addressed in 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; and 4. Integrate enhanced H_2ase mutants and culture and metabolic factor studies to maximize 24-hour H_2 production.

This poster describes a screening assay, compatible with high-throughput bioprospecting or molecular biology methods, for assessing biological hydrogen (H_2) production. While the assay is adaptable to various physical configurations, we describe its use in a 96-well, microtiter plate format with a lower plate containing H_2-producing cyanobacteria strains and controls, and an upper, membrane-bottom plate containing a color indicator and a catalyst. H_2 produced by cells in the lower plate diffuses through the membrane into the upper plate, causing a color change that can be quantified with a microplate reader. We used response surface methodology to optimize the concentrations of the components in the upper plate. The assay is reproducible, semi-quantitative, sensitive to 20 nmol of H_2 or less, and largely unaffected by oxygen, carbon dioxide, or volatile fatty acids at levels appropriate to biological systems.

24

GTL

Hydrogen Production by PSS1, A *hoxH* Mutant of *Synechocystis sp.* PCC 6803

Hatem Mohamed,[1*] Paul S. Schrader,[2,3] Elizabeth H. Burrows,[1] and **Roger L. Ely**[1,2] (ely@engr.orst.edu)

[1]Department of Biological and Ecological Engineering, Oregon State University, Corvallis, Oregon; [2]Department of Chemical Engineering, Yale University, New Haven, Connecticut; and [3]Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon

Project Goals: See goals for abstract 23.

In this research we created a *hoxH* mutant of *Synechocystis* sp. PCC 6803, identified as PSS1, by drastically modifying a portion of the enzyme predicted to be adjacent to the active site. Sequence analyses of the genomes of *Ralstonia eutropha*, *Desulfovibrio gigas*, and *Synechocystis* sp. PCC 6803 show five conserved regions in the *hoxH* gene, designated as L1 through L5. By deleting 300 base pairs of the gene, we removed all of region L5 and part of L4. This poster describes hydrogen (H_2) production performance of the mutant, with and without inhibitors of specific metabolic pathways, in high-throughput screening assay studies, GC vial tests, photobioreactor studies, and membrane inlet mass spectrometer measurements. In some ways, even with two of the five conserved regions eliminated, PSS1 behaves quite similarly to wild-type (WT) PCC 6803. However, some measurement methods show greater H_2 production from the mutant than from WT cells, while others show no H_2 production at all.

25

MEWG

Metabolic Flux Analysis of Metabolism in *Synechocystis* sp. *PCC 6803* for Improving Hydrogen Production

Frank W.R. Chaplen,* Elizabeth H. Burrows, and **Roger L. Ely** (ely@engr.orst.edu)

Department of Biological and Ecological Engineering, Oregon State University, Corvallis, Oregon

Project Goals: See goals for abstract 23.

* Presenting author

Metabolic engineering of *Synechocystis* sp. PCC 6803 strains with the capability of consistent, high-yield, biosolar hydrogen (H_2) production requires continued development of comprehensive mathematical models describing the metabolism underlying H_2 production and linking genomic, proteomic, and metabolomic information. Such models help to organize disparate hierarchical information, discover new strategies, and understand the essential qualitative features of components and interactions in a complex system[1]. Metabolic flux analysis is an analytical approach used to estimate the fluxes through a biochemical reaction network operating at steady state based on measured inflows and outflows. Analysis of both over- and underdetermined networks is possible, the latter with linear programming. Here we use metabolic flux analysis to examine the effect of different network parameters and constraints on photoautotrophic H_2 production by wild type (WT) *Synechocystis* sp. PCC 6803 and by a high H_2-producing mutant (M55) with impaired Type I NADH-dehydrogenase (NDH-1) function. Two different networks are used with both WT and M55 mutant strains under chemostat growth: 1) an overdetermined network with 24 metabolites and 20 constraints, requiring at least 4 measurements of fermentation parameters for solution; and 2) an underdetermined network with increased detail for gene knockout simulations, requiring constraints-based approaches for solution1. The inflows and outflows measured for both networks are H_2, O_2, CO_2, glucose, glycogen, ammonium, and biomass production/consumption. The behavior of both model networks is consistent with WT and M55 mutant phenotypes, thus validating the general approach. The models are then used to provide insights into the possible effects of different mutant phenotypes on H_2 production.

References

1. Bailey, J.E. (1998) Mathematical modeling and analysis in biochemical engineering: Past accomplishments and future opportunities. Biotechnol. Prog. 14:8-20.
2. Price, N.D., Reed, J.L. and Palsson, B. Ø. (2004) Genome-Scale Models of Microbial cells: Evaluating the Consequences of Constraints. Nature Review|Microbiology 2:888-897.

26

GTL

Optimization of Media Nutrient Composition for Increased Photofermentative Hydrogen Production by *Synechocystis* sp. PCC 6803

Elizabeth H. Burrows,* Frank W.R. Chaplen, and **Roger L. Ely** (ely@engr.orst.edu)

Department of Biological and Ecological Engineering, Oregon State University, Corvallis, Oregon

Project Goals: See goals for abstract 23.

By optimizing concentrations of key components in nutrient media, we achieved over 60-fold greater photofermentative hydrogen (H_2) production by *Synechocystis* sp. PCC 6803 than was achieved by analogous, sulfur-deprived cultures, which produce more H_2 than cultures grown on complete media. We used response surface methodology to determine optimum conditions and found that, instead of completely starving cells of sulfur or nitrogen, the highest H_2 production occurred with low concentrations of S and N. H_2 profiling experiments provided initial screening of NH_4^+, HCO_3^-, SO_4^{2-}, and PO_4^{3-} concentrations and identified the significant variables for H_2 production to be NH_4^+, SO_4^{2-}, and the interactions of both NH_4^+ and SO_4^{2-} with HCO_3^-. A central composite design was implemented and subsequent response surface analysis of the data resulted in a saddle point. Ridge analysis was then conducted to identify high points within the region of interest, and those concentrations were tested and compared with sulfur-deprived cells. Our results indicate that optimized amounts of nitrogen and sulfur in the nutrient media are superior to total deprivation of these nutrients for H_2 production.

27

GTL

Performance of REHX, A *Synechocystis* sp. PCC 6803 Mutant with an Oxygen-Tolerant hoxH Subunit from *Ralstonia eutropha*

Hatem Mohamed,*[1] Paul S. Schrader,[2,3] Elizabeth H. Burrows,[1] and **Roger L. Ely**[1,2] (ely@engr.orst.edu)

[1]Department of Biological and Ecological Engineering, Oregon State University, Corvallis, Oregon;
[2]Department of Chemical Engineering, Yale University,

* Presenting author

New Haven, Connecticut; and [3]Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon

Project Goals: See goals for abstract 23.

This poster describes hydrogen (H_2) production behavior of a *Synechocystis* sp. PCC 6803 mutant strain, designated as REHX. In REHX, we replaced the gene encoding the wild-type (WT) PCC 6803 hoxH subunit with the *hoxH* gene from the soluble hydrogenase of *Ralstonia eutropha,* which has been well characterized, is oxygen tolerant, and shares significant sequence homology with the PCC 6803 *hoxH*. We also inserted *hypX*, a gene that encodes an accessory protein essential for oxygen tolerance in *R. eutropha*. The *R. eutropha hoxH* and *hypX* genes were both transcribed into mRNA and hydrogen production was observed, implying that the foreign *hoxH* gene was translated and that the protein functioned with the unaltered subunits of the WT hydrogenase. H_2 production and oxygen tolerance of REHX were evaluated, with and without inhibitors of specific metabolic pathways, in high-throughput screening assay studies and in gas chromatograph and membrane inlet mass spectrometer measurements.

28

GTL

Understanding and Engineering Electron Flow to Nitrogenase to Improve Hydrogen Production by Photosynthetic *Rhodopseudomonas palustris*

James 'Jake' McKinlay* (mckinla1@u.washington.edu), Erin Heiniger, Jean Huang, and **Caroline S. Harwood** (csh5@u.washington.edu)

Department of Microbiology, University of Washington, Seattle, Washington

Project Goals: To develop and apply techniques in metabolic engineering to improve the biocatalytic potential of the bacterium *Rhodopseudomonas palustris* for nitrogenase-catalyzed hydrogen production. *R. palustris*, is an ideal platform to develop as a biocatalyst for hydrogen gas production because it is an extremely versatile microbe that produces copious amounts of hydrogen by drawing on abundant natural resources of sunlight and biomass. Anoxygenic photosynthetic bacteria, such as *R. palustris*, generate hydrogen and ammonia during a process known as biological nitrogen fixation. This reaction is catalyzed by the enzyme nitrogenase and normally consumes nitrogen gas, ATP and electrons. The applied use of nitrogenase for hydrogen production is attractive because hydrogen is an obligatory product of this enzyme and is formed as the only product when nitrogen gas is not supplied. Our challenge is to understand the systems biology of *R. palustris* sufficiently well to be able to engineer cells to produce hydrogen continuously, as fast as possible and with as high a conversion efficiency as possible of light and electron donating substrates.

Rising energy demands and the imperative to reduce carbon dioxide emissions are stimulating research on the development of bio-based fuels. Hydrogen gas is one of the most promising biofuels, having about three times the energy content of gasoline. The purple bacterium *Rhodopseudomonas palustris* naturally uses energy from sunlight and electrons from organic waste to produce H_2 and ammonia using any of its three nitrogenases (1). We have also described *R. palustris* mutants that express nitrogenase at all times and use this enzyme to produce pure H_2 without accompanying ammonia production (2). Although we achieved higher specific H_2 productivities using these mutants, H_2 yield and rate can still be improved further. In order to effectively engineer *R. palustris* for improved H_2 production, we need to better understand the pathways and proteins that are used by cells to transfer electrons from electron donating substrates (e.g., acetate) to nitrogenase; the site of H_2 production.

To identify which metabolic reactions provide electrons for H_2 production we are conducting ^{13}C-labelling experiments to compare carbon flux distributions between non-H_2-producing wild type and constitutively H_2-producing mutant strains. We are also using genetic and biochemical approaches to identify electron carriers that operate between central metabolism and nitrogenase. Mutational analysis has indicated that the FixABCX complex transports electrons to nitrogenase but also that it is not the sole electron carrier involved. We are therefore also investigating novel candidate electron carriers that were up-regulated in mutants with constitutive nitrogenase activity (2).

Once potential bottlenecks in electron transfer are identified, various genetic tools will be used to engineer *R. palustris* for improved H_2 production. Towards this end we have been working to develop a plasmid vector for controlled gene expression. Recently, members of our group identified a new signaling compound from *R. palustris* that acts with the transcriptional regulator, RpaR, to control expression of a gene named *rpaI* (Schaefer et al. submitted). We have developed an expres-

* Presenting author

sion vector based on this novel signaling system whereby varying the amount of signaling molecule provided as the inducer can control gene expression levels over a dynamic range. We anticipate that we will be able to combine new information about mechanisms of electron flow to nitrogenase with the use of our new gene expression system and other tools to improve rates of H_2 production by *R. palustris*.

References

1. Oda, Y., S. K. Samanta, F. E. Rey, L. Wu, X. Liu, T. Yan, J. Zhou, and C. S. Harwood. 2005. Functional genomic analysis of three nitrogenase isozymes in the photosynthetic bacterium *Rhodopseudomonas palustris*. *J Bacteriol* 187:7784-94.
2. Rey, F. E., E. K. Heiniger, and C. S. Harwood. 2007. Redirection of metabolism for biological hydrogen production. *Appl Environ Microbiol* 73:1665-71.

29

GTL

Modeling Electron Flow in *Rhodobacter sphaeroides* to Quantitatively Identify Approaches to Maximize Hydrogen Production

L. Safak Yilmaz[1,*] (yilmaz@cae.wisc.edu), Yi-Kai Su,[1] Wayne S. Kontur,[2] Timothy J. Donohue[2] (tdonohue@bact.wisc.edu), and **Daniel R. Noguera[1]** (noguera@engr.wisc.edu)

[1]Department of Civil and Environmental Engineering and [2]Department of Bacteriology, University of Wisconsin, Madison, Wisconsin

Project Goals: To determine the impact of metabolic and regulatory networks on light-powered hydrogen production by *Rhodobacter sphaeroides*. Our experiments will determine the quantitative contribution of networks that are known or predicted to impact solar-powered hydrogen production. We will use genetic, genomic, and modeling approaches to quantify the role of previously uncharacterized networks, ultimately using this information to design microbial bioreactors to efficiently produce hydrogen from solar energy.

This is a new project aimed at quantifying electron flow within networks that impact solar-powered H_2 production by the photosynthetic bacterium *Rhodobacter sphaeroides*. We plan to use genetic, genomic, and modeling approaches to quantify the role of known and previously uncharacterized networks in solar-powered H_2 production by this organism. Ultimately, this information can be used to design microbial bioreactors that efficiently generate H_2 or other alternative fuels from solar energy.

The initial work has centered on the development of a mathematical model to represent the electron flow from an organic substrate to the different and competing electron accepting reactions, during anaerobic photosynthetic growth. As a preliminary step, the model describes the quantitative apportionment of electrons derived from the oxidation of an organic substrate to pathways related to cell mass synthesis, polyhydroxyalkanoate (PHA) formation, hydrogen production, and accumulation of soluble organic products. Experimental measurements of substrate, biomass accumulation, nitrogen and phosphorus consumption, and pH are used to calculate best-fit estimates of the fraction of electrons consumed in each pathway.

When the model is fit to data from batch experiments with exponentially growing *R. sphaeroides* 2.4.1. and succinate as the sole organic substrate, it predicts that 47% of the electron flow is related to biomass formation, 41% to PHA accumulation, and 12% to soluble microbial products. In experiments with stationary phase cultures with nitrogen limitation, 26% of the electrons are predicted to be used in PHA synthesis, 26% in hydrogen production, and 48% in the formation of reduced soluble organic substrates. When propionate is used as the sole organic substrate, the predicted apportionment of electrons during exponential growth is similar to that seen with succinate, but under nitrogen limiting conditions, the model predicts lower PHA formation, higher hydrogen production, and a higher formation of soluble organic substrates.

These preliminary evaluations are being used to inform genetic strategies to optimize hydrogen production. Finally, as more complete experimental methods are developed, we expect to be able to experimentally validate the model predictions regarding PHA formation and hydrogen production, as well as to identify the type of soluble organic substrates produced by *R. sphaeroides*.

* Presenting author

30

GTL

Towards Experimental Verification of Protein Coding Transcripts of *Chlamydomonas reinhardtii* and Comprehensive Modeling of its Metabolic Network

Ryan Murray,[1] Chenwei Lin,[1] Arvind Chavali,[2] Erik F.Y. Hom,[3] Xinping Yang,[1] David E. Hill,[1] Marc Vidal,[1] Jason Papin,[2] and **Kourosh Salehi-Ashtiani**[1]* (Kourosh_Salehi-ashtiani@DFCI.harvard.edu)

[1]Center for Cancer Systems Biology (CCSB), Department of Cancer Biology, Dana-Farber Cancer Institute, and Department of Genetics, Harvard Medical School, Boston, Massachusetts; [2]Department of Biomedical Engineering, University of Virginia, Charlottesville, Virginia; and [3]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts

Project Goals: Our objectives for this project are to experimentally verify, define, and validate metabolic protein-coding genes of *Chlamydomonas reinhardtii* and model a comprehensive metabolic network for this organism. The proposed experiments utilize a technology platform that can be adapted to virtually any organism, and hence serve as a prototype that can be used for gene validation in any species. *Chlamydomonas reinhardtii* is an ideal organism for this project because 1) it is an important "bio-energy" organism, and 2) a draft of its genome sequence is currently available. The obtained results will be used to build a more complete model of the metabolic circuitry of this organism. The generation of a metabolic network will in turn help validate examined genes by defining a biological role for them. From our obtained results, we should be able to formulate testable hypotheses as to how to optimize bio-fuel (including hydrogen gas) production in this organism. To achieve these objectives, we will be carrying out experiments to define and verify transcript structures of metabolic genes in *Chlamydomonas reinhardtii* by RT-PCR and RACE, functionally validate the transcripts by yeast two-hybrid experiments, and build and interpret predictive metabolic network models based on the obtained results.

Chlamydomonas reinhardtii is a promising "bio-energy" organism capable of producing hydrogen gas and other "bio-fuel" resources. Although a draft of its genome is available, annotations of most of its protein coding genes have not been experimentally verified. Furthermore, a comprehensive metabolic map has not been generated for this organism.

To experimentally verify, define, and validate metabolically related protein-coding genes of *C. reinhardtii*, and to produce a comprehensive metabolic map, focusing on hydrogen production and other bio-energy aspects, we plan to: *i*) experimentally verify and define the transcript structures of ~2,000 metabolically related genes and clone their open reading frames (ORF) for downstream protein-based validation studies; *ii*) identify protein-protein interactions among the metabolic gene products; and *iii*) build protein interaction maps and metabolic networks that will ultimately allow the development of testable predictions of *C. reinhardtii* physiology, including gene lethality and rates of growth under defined environmental conditions.

Here we present our computational and experimental results towards completing the aforementioned goals. We have computationally examined the current JGI *C. reinhardtii* annotated transcripts (frozen version 3.1 release) for the presence of full length ORFs. Our computational analyses indicate that of the ~14,500 annotated transcripts, ~9,500 appear to be "complete", *i.e.* they each appear to contain a complete ORF, while the remaining ~5,000 appear to be "incomplete". Starting from KEGG and KOG databases, we selected a list of ~470 transcripts for experimental verification by RT-PCR and RACE. Our RT-PCR experiments on a large subset of these showed that the majority of transcripts with a predicted full length ORF could indeed be amplified by PCR. Cloning and sequencing of these transcripts are in progress to further verify these transcript models.

Using currently available resources, we have generated a draft for the intracellular central metabolic network of *Chlamydomonas reinhardtii*. Our *in silico* model accounts for key pathways involved in carbohydrate and energy metabolism, including glycolysis, TCA cycle, pentose phosphate pathway and oxidative phosphorylation. The current network accounts for 224 *gene products* (214 nuclearly- and 7 mitochondrially-encoded), 96 reactions, and 140 metabolites, partitioned into 6 sub-compartments: chloroplast, mitochondria, flagella, cytosol, extracellular space, and glyoxysome. Approximately 82% of the network reactions were based on gene predictions/annotations, while the remainder were inferred to ensure model completeness, mainly to account for metabolite exchange between compartments. Our metabolic model suggests a number of adjustments are needed in the current *C. reinhardtii* gene annotations.

* Presenting author

Our iterative approach of integrating experimental data with model-building enables us to refine genome annotations and metabolic network maps progressively. We are currently in the process of completing the first of such cycles. Concurrent with carrying on these iterations, we will develop testable hypotheses to accelerate future biological discoveries relevant to our goals and those of the *C. reinhardtii* scientific community at large.

31

GTL

Addressing Unknown Constants and Metabolic Network Behaviors through Petascale Computing: Understanding H_2 Production in Green Algae

Christopher H. Chang,[1]* David Alber,[1] Peter Graf,[1] Kwiseon Kim,[1] Arthur R. Grossman,[2] Glenn Murray,[3] Matthew C. Posewitz,[3] and **Michael Seibert**[1] (mike_seibert@nrel.gov)

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

Project Goals: Development of photobiological H_2-production processes, a key component in pursuit of DOE's renewable energy mission, would be substantially accelerated by increasing our understanding of the extremely complex underlying biology. To address biological complexity at the molecular level, systems biology has evolved rapidly in recent years. However, the lack of comprehensive experimental data for a given organism prevents reliable predictive modeling. We will therefore employ petascale computing to address this issue by computational parameter estimation to delimit the space of stable solutions for experimentally constrained metabolic models. The response will be characterized at the level of enzyme kinetic differential equation parameters. Through development of scalable software tools, iterative model building, and incorporation of experimental constraints generated by high-throughput "omics" technologies, a model of metabolism linked to H_2 production in the green alga, *Chlamydomonas reinhardtii*, will be constructed. Once a set of acceptable kinetic parameters have been computed, the model will then be used for high performance optimization of H_2 output in the space of enzyme expression levels, subject to limitations on cell viability. Integration into the popular Systems Biology Workbench will make our tools accessible to the general user community. The work is envisioned as an important contribution toward long-term development of a complete in silico cell.

The Genomics revolution has resulted in a massive and growing quantity of whole-genome DNA sequences, which encode the metabolic catalysts and ribonucleic acids necessary for life. However, gene annotations can rarely be complete, and measurement of the kinetic constants associated with the encoded enzymes can not possibly keep pace, necessitating the use of careful modeling to explore plausible network behaviors. Key challenges are (a) the quantitative formulation of the kinetic laws governing each transformation in a fixed model network; (b) characterizing the stable solutions of the associated ordinary differential equations; (c) fitting the latter to metabolomics data as it becomes available; and (d) optimizing a model output against the possible space of kinetic parameters, with respect to properties such as robustness of network response or maximum consumption/production. This project addresses this large-scale uncertainty in the genome-scale metabolic network of the water-splitting, H_2-producing green alga, *Chlamydomonas reinhardtii* [1,2]. Each metabolic transformation is formulated as a steady-state process in such manner that the vast literature on known enzyme mechanisms may be incorporated directly. We have encoded glycolysis, the tricarboxylic acid cycle, and basic fermentation pathways in Systems Biology Markup Language (SBML) with careful annotation and consistency with the KEGG database, yielding a preliminary model with 4 compartments, 85 species, 35 reactions, and 89 kinetic constants.

The SemanticSBML toolkit is first used to combine, validate, and annotate hand-coded SBML models corresponding to metabolic modules. In this manner, a genome-scale kinetic model may be constructed from a library of simple component pathways, making modification and maintenance plausible. The use of a standard, XML-based language and preservation of semantics in the naming convention for variables and parameters means that customization of the library to any organism may be accomplished easily. From an instance of a model, use of libSBML and SOSLib allows automatic production of a C program that when executed, optimizes the model's kinetic parameters according arbitrary test criteria. The generation of this optimizer from the model consists of several steps. First, the unified SBML model is parsed and converted to a system of ordinary differential equations (ODEs), including Jacobian and sensitivity matrices. These are then translated to C functions and embedded in code utilizing the ODE solver package CVODES, resulting in a library that can efficiently simulate the model, including calculating derivatives with

* Presenting author

respect to parameters (in our case enzyme kinetic constants). This library is in turn incorporated into code that calculates the objective functions implied by challenges (2) – (4) above, as well as the objective function derivatives. Finally these routines are built into code using the optimization package TAO to optimize the model with respect to the kinetic parameters. We illustrate the system and present numerical results.

Future software development will include a parallel global optimization algorithm, enhanced integration with the Systems Biology Workbench, and development of distributed data analytics and visualization tools, a graphical interface, and interactive visualization for high-dimensional datasets. The free availability of these tools will allow the broader biological community to sample and/or optimize genome-scale metabolic networks in the space of thousands of parameters, and permit study of, for example, implications of enzymatic co-evolution and system-wide metabolic engineering. The software design is specifically targeting the combined goals of no-cost availability, open standards, and wide accessibility to researchers from all sectors. The size of the addressable problems will steadily increase with the availability of parallel computing platforms ranging over now-common desktop multi-core/multi-chip workstations, massive distributed grids, workgroup clusters, and cutting-edge petascale architectures.

References

1. Mus, F., A. Dubini, M. Seibert, M.C. Posewitz and A.R. Grossman. (2007) "Anaerobic acclimation in *Chlamydomonas reinhardtii*: Anoxic gene expression, hydrogenase induction and metabolic pathways", *J. Biol. Chem.* **282 (35)**, 25475-25486.
2. Chang, C., D. Alber, P.Graf, K. Kim, and M. Seibert (2007) "Addressing unknown constants and metabolic network behaviors through petascale computing: understanding H_2 production in green algae," *J. Phys: Conf. Ser.* **78**: 012011.

32

GTL

Flexibility of Algal Anaerobic Metabolism is Revealed in a Mutant of *Chlamydomonas reinhardtii* Lacking Hydrogenase Activity

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

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

Project Goals: Past research has shown that photosynthesis, respiration, and fermentation are all required to sustain H_2 photoproduction from water in algae. These microbes utilize [FeFe]-hydrogenases, which are the most efficient H_2-generating biocatalysts known. The long-term objective of our project is to identify the suite of genes facilitating and/or limiting H_2 photoproduction in the alga, *Chlamydomonas reinhardtii*, by conducting global gene expression and cell metabolism studies using algal cells acclimated to conditions known to induce H_2-production activity. A detailed understanding of the influences of metabolism and other environmental factors on the coordinated expression of genes and biochemical pathways associated with H_2-production activity will ultimately be required to increase the yields of renewable H_2 production for potential future applications. To accomplish this we will examine WT cells and a number of NRELs H_2-production mutants under a number of experimental conditions using *Chlamydomonas* gene microarrays along with extensive biochemical assays. Algal H_2 production requires the synergies of multiple redox proteins, sensors, biochemical pathways and regulatory processes. Knowledge gained by deconvoluting these interactions will help us identify specific targets for future strain engineering aimed at enhancing H_2 production in *C. reinhardtii*.

The green alga, *Chlamydomonas reinhardtii,* has an extensive network of fermentation pathways that are activated when the cells acclimate to anoxia, and hydrogenase activity is an important component of this metabolism. *Chlamydomonas* uses fermentative pathways for ATP production during anoxia in the dark, catabolising starch into the predominant fermentation products formate, acetate, ethanol, CO_2, and H_2 in what is classified as heterofermentation. Previous microarray studies as well as RT-PCR analysis identified increases in specific gene

* Presenting author

transcripts under these conditions [1]. Indeed, anoxia leads to the up-regulation of genes encoding proteins involved in fermentation and more specifically those associated with pyruvate catabolism, such as pyruvate formate lyase (PFL1) and pyruvate:ferredoxin oxidoreductase (PRF1). Moreover, increased levels of transcripts encoding several regulatory elements suggest that activation of specific signalling pathways and the need to control translational and post-translational processes occur in the cells as the environment becomes anoxic.

In this study we have compared metabolic and regulatory acclimation responses that accompany anaerobiosis in wild-type cells and in a mutant defective for H_2 production as a consequence of a null mutation in the [FeFe]-hydrogenase maturation protein, HYDEF [2]. The mutant exhibits both elevated accumulation of succinate and diminished production of CO_2, relative to the parental strain, after four hours of dark, anaerobic acclimation. These results are consistent with increased activity of enzymes required for anoxic succinate production, and a decreased metabolic flux through the PRF1 pathway. In the absence of hydrogenase activity, an increase in succinate suggests the need to activate alternative pathways to metabolize pyruvate and re-oxidize NAD(P)H, which allows continued glycolysis and fermentation in the absence of O_2. Activities required for succinate production include pyruvate carboxylation and/or oxaloacetate reduction, which generate malate. Malate can then be further metabolized to fumarate and finally to succinate. Enzymes that can potentially catalyze the carboxylation of pyruvate to malate *via* independent pathways are the pyruvate carboxylase and malic enzymes. Marked increases in the abundance of mRNAs encoding both of these enzymes are observed in the mutant relative to the parental strain. *Chlamydomonas* has a single gene encoding pyruvate carboxylase and six genes encoding putative malic enzymes. Only one of the malic enzyme genes, *MME4*, shows a dramatic increase in expression (mRNA abundance) in the *hydEF-1* mutant during anaerobiosis. Furthermore, there are also large increases in transcripts encoding fumarase and fumarate reductase, the enzymes required for the conversion of malate to succinate. To further identify potential metabolic and regulatory features of the *hydEF-1* mutant relative to the parental strain, we used a high density, oligonucleotide (70-mer)-based microarray to compare genome-wide transcript patterns of *hydEF-1* and parental strains after transferring cultures from aerobic to anaerobic conditions. Several transcripts encoding proteins associated with cellular redox functions and other aspects of anaerobic metabolism were observed to be differentially regulated in the mutant under anaerobic conditions.

In summary, these experiments illustrate the marked metabolic flexibility of *Chlamydomonas* and also provide insights into how mutants, altered in normal H_2 metabolism, acclimate to H_2-producting conditions. Moreover the information obtained will provide the foundation for metabolic engineering and accurate *in silico* modelling of *Chlamydomonas* metabolism. The availability of the *Chlamydomonas* genome sequence, combined with high-throughput-'omics'-based approaches is critical in this effort. Moreover, the use of specific mutant strains can help establish the foundation for a more comprehensive understanding of metabolic networks, how cells adjust metabolite fluxes when specific metabolic reactions are blocked, and how to improve H_2-production yields.

References

1. Mus, F., A. Dubini, M. Seibert, M.C. Posewitz and A.R. Grossman. (2007) "Anaerobic acclimation in *Chlamydomonas reinhardtii*: Anoxic gene expression, hydrogenase induction and metabolic pathways", *J. Biol. Chem.* **282 (35)**, 25475-25486.
2. Posewitz, M. C., P. W. King, S. L. Smolinski, L. Zhang, M. Seibert, and Maria L. Ghirardi (2004) "Discovery of Two Novel Radical SAM Proteins Required for the Assembly of an Active [Fe]-Hydrogenase," *J. Biol. Chem.* **279**, 25711-25720.

33

GTL

Development of Biologically Based Assays to Study Rate-Limiting Factors in Algal Hydrogen Photoproduction

Alexandra Dubini, Matt Wecker, and **Maria L. Ghirardi*** (maria_ghirardi@nrel.gov)

National Renewable Energy Laboratory, Golden, Colorado

Photobiological H_2 production from water is a clean, non-polluting and renewable technology. Although the potential light conversion efficiency to H_2 by biological organisms is theoretically high (about 10%), the system is currently limited by biochemical and engineering constraints. These limitations include (but are not restricted to) the extreme O_2 sensitivity of the biological hydrogenases and the low availability of reductants to the hydrogenase due to the existence of competing metabolic pathways. Our research addresses the O_2 sensitivity issue by developing a new, biologically-based assay to screen large microbial populations for improved H_2-production properties. This novel assay is based on the H_2-sensing properties of systems found in nitrogenase-containing

* Presenting author

photosynthetic bacteria. We will validate the new assay by using it to screen mutants generated through directed-evolution techniques for O_2 tolerant [FeFe]-hydrogenases. To address the issue of competitive metabolic pathways with H_2 production, we will adapt the yeast two-hybrid assay to measure the interactions between different ferredoxin isoforms present in *Chlamydomonas reinhardtii* with different proteins known to accept electrons from ferredoxin in most photosynthetic and fermentative organisms. An expanded approach will include the use of ferredoxin probes to identify unknown target genes out of a *C. reinhardtii* expression library. Identified protein-protein interactions above will be quantified by isothermal titration calorimetry. The information obtained will guide future protein and metabolic engineering efforts to divert most of the electron flux from ferredoxin to the hydrogenase. This work will develop techniques that will drive a deeper understanding of algal H_2 metabolism and accelerate the development of future photobiological H_2-production catalysts and organisms.

34

GTL

Quantitative Tools for Dissection of Hydrogen-Producing Metabolic Networks

Joshua Rabinowitz* (joshr@genomics.princeton.edu), G. Charles Dismukes, and Herschel Rabitz

Department of Chemistry and Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey

Project Goals: With an eye towards eventually enabling rational optimization of microbial hydrogen production, here we aim to develop novel tools for quantitative dissection of hydrogen-producing metabolic networks. These tools will bring innovations both in high-throughput experimental metabolomics and in algorithms for predictive modeling/analysis of experiments. We will first develop quantitative experimental tools for simultaneous measurements of multiple metabolite concentrations and fluxes. These will include liquid chromatography-tandem mass spectrometry (LC-MS/MS) assays and complementary electrochemical, fluorescence, or NMR assays. We will then apply these tools to map the dynamic metabolic responses of *Clostridium acetobutylicum* and *Synechococcus* species to changing environmental conditions. Finally, we will use the data to guide the development of predictive models of hydrogen metabolism. The development of predictive metabolic models will pave the way to computationally-guided optimization of microbial hydrogen yields.

Overview: With an eye towards eventually enabling rational optimization of microbial H_2 production, we have recently initiated a project that aims to develop novel tools for quantitative dissection of H_2-producing metabolic networks. We hope that these tools will bring innovations both in high-throughput experimental metabolomics and in algorithms for predictive modeling/analysis of experiments. The project has three major objectives:

Aim 1: Develop quantitative experimental tools for simultaneous measurements of multiple metabolite concentrations and fluxes. We aim to create and apply a comprehensive set of experimental tools for measuring multiple (200+) intermediate metabolites and gases (including H_2) implicated in energy metabolism. Metabolomic data will be generated by liquid chromatography-tandem mass spectrometry (LC-MS/MS). We have developed methods that enable quantitation of several hundred metabolites of general interest by LC-MS/MS, and are working to expand to include also compounds of specific importance to hydrogen metabolism. For compounds of particular importance that cannot be reliably measured by LC-MS/MS, complementary electrochemical, fluorescence, or NMR assays will be employed. A particular interest of our group is development also of highly sensitive devices for electrochemical monitoring of dissolved O_2 or H_2 that enable simultaneous measurement intracellular levels of NAD(P)H. Exemplary LC-MS/MS data will be presented, as will be data regarding monitoring of dissolved H_2 and intracellular NAD(P)H.

Aim 2: Map the organisms' dynamic metabolic responses to changing environmental conditions. Metabolic network activity, including H_2 production, is highly sensitive to the availability of a broad range of nutrients and environmental stresses. To elucidate the interactions of these inputs, we plan to modulate the cellular environment and apply the measurement techniques developed in Aim 1 to acquire high quality dynamic data describing the full network response. Currently, we are working on optimizing cell handling for *Clostridium* and *Synechococcus* species, the H_2-producing organisms of greatest interest to us. Given the preliminary status of our work on these organisms, we tentatively plan to present here exemplary data from *E. coli* to highlight the potential of the approach.

* Presenting author

Aim 3: Develop predictive models of H_2-metabolism. The emerging ability to generate high quality, dynamic metabolomic data raises a new challenge: how to apply these data to (a) better understand metabolic regulation, and (b) develop predictive, quantitative metabolic models. We plan to achieve these objectives via integrated computational-experimental approaches, which use experimental data to drive computational modeling, while also using computational results to guide experimental design. An example of such an effort related to nitrogen metabolism in *E. coli*, although not directly relevant to hydrogen production, will be presented to highlight the concepts involved. We are also eager to discuss our computational plans in the hydrogen arena with groups conducting related research.

Organisms: The foci of the present newly initiated project are *Clostridium acetobutylicum* (possessing the fastest and highest yielding hexose fermentation pathway to H_2 of any microbe yet reported), and a new class of thermophilic cyanobacteria that lack hydrogenase genes and produce H_2 at 62°C via a nitrogenase-dependent pathway (*Synechococcus* species). We are eager to build contacts with experts in these organisms.

35

GTL

Quantitative Tools for Characterization of H_2-Producing Cyanobacteria

Gennady Ananyev, Nicholas Skizim, and **Charles Dismukes*** (dismukes@Princeton.edu)

Department of Chemistry and Environmental Institute, Princeton University, Princeton, New Jersey

This subproject of DE-FG02-07ER64488 focuses on quantitative tools for biohydrogen detection and applications to N_2-fixing cyanobacteria. We have built a cell for simultaneous real-time detection of extracellular dissolved H_2 and intracellular reduced pyridine nucleotide levels. Ultrasensitive H_2 detection (>1 nanomolar) is accomplished with a homebuilt Clark-type electrochemical cell using live cells (3 - 5 μl volume at 1 - 8 μg dry wt. equiv.). NADPH+NADH concentrations are assayed using wavelength selective UV light-emitting diode (360 nm) excitation and fluorescence emission (450 +/- 20 nm) detection. These tools have revealed for the first time two major temporal phases of H_2 production in anaerobically poised cyanobacteria arising from reductant pools that equilibrate with [NiFe]-hydrogenase. The first phase correlates with the availability of residual NADPH produced in prior photosynthetic stage, while phase 2 arises from NADH produced by anaerobic fermentation of intracellular glycogen. Experiments with N_2-fixing cyanobacteria reveal additional kinetic phases of H_2 production presumed due to interactions with N_2-fixation pathways and which correlate with the cells' circadian rhythms. We shall describe how these tools have been used to identify more O_2-tolerant H_2-producing cyanobacteria isolated during bioprospecting studies.

* Presenting author

Systems Environmental Microbiology

36

GTL

The Virtual Institute of Microbial Stress and Survival VIMSS:ESPP Overview

Adam P. Arkin[1,2,4,5*] (aparkin@lbl.gov), **Terry C. Hazen**,[1,2*] Carl Abulencia,[1,7] **Eric J. Alm**,[1,3] Gary Anderson,[1,2] Manfred Auer,[1,2] Edward Baidoo,[1,2] **Kelly S. Bender**,[1,12] Peter Benke,[1,2] Sharon Borglin,[1,2] Eoin Brodie,[1,2] Steven Brown,[1,6] Lauren Camp,[1,2] Romy Chakraborty,[1,2] Swapnil Chhabra,[1,2] Gabriela Chirica,[1,10] Dylan Chivian,[1,2] Michael Cipriano,[1,2] Paramvir S. Dehal,[1,2] Todd DeSantis,[1,2] Elliot Drury,[1,8] **Inna Dubchak**,[1,2] Dwayne Elias,[1,8] **Matthew W. Fields**,[1,10] Veronica On Yi Fok,[1,2] Julian Fortney,[1,2] Sara Gaucher,[1,13] Jil Geller,[1,2] Masood Hadi,[1,13] Zhili He,[1,11] Chris Hemme,[1,13] Kristina Hillesland,[1,9] Hoi-Ying Holman,[1,2] Katherine H. Huang,[1,2] Y. Wayne Huang,[1,2] Chiachi Hwang,[1,10] Janet Jacobsen,[1,2] Marcin P. Joachimiak,[1,2] Dominique Joyner,[1,2] **Jay Keasling**,[1,2,4] Keith Keller,[1,2] Kimberly L. Keller,[1,8] **Martin Keller**,[1,6] Yooli Light,[1,13] Eric Luning,[1,2] Robert Meagher,[1,13] **Aindrila Mukhopadhyay**,[1,2] Anthony Palumbo,[1,6] Richard Phan,[1,2] Tommy Phelps,[1,6] Francesco Pingitore,[1,2] Mircea Podar,[1,6] Morgan N. Price,[1,2] Alyssa Redding,[1,2] Jarrod Robertson,[1,8] Rajat Sapra,[1,13] Christopher Schadt,[1,6] Matthew Shirley,[1,8] Amy Shutkin,[1,2] Mary Singer,[1,2] **Anup Singh**,[1,13] **David A. Stahl**,[1,9] Sergey Stolyar,[1,9] Anitha Sundararajan,[1,10] Yinjie Tang,[1,2] Joy Van Nostrand,[1,11] Sandra Villa,[1,2] Christopher Walker,[1,9] **Judy D. Wall**,[1,8] Zamin Koo Yang,[1,6] Huei-che Yen,[1,8] Grant Zane,[1,8] Aifen Zhou,[1,11] and **Jizhong Zhou**[1,11]

[1]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov/; [2]Lawrence Berkeley National Laboratory, Berkeley, California; [3]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts; [4]University of California, Berkeley, California; [5]Department of Bioengineering, University of California, Berkeley, California; [6]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [7]Diversa, Inc., San Diego, California; [8]University of Missouri, Columbia, Missouri; [9]University of Washington, Seattle, Washington; [10]Montana State University, Bozeman, Washington; [11]University of Oklahoma, Norman, Oklahoma; [12]Southern Illinois University; and [13]Sandia National Laboratories, Livermore, California

Project Goals (Abstracts 36-62): The Virtual Institute of Microbial Stress and Survival (http://vimss.lbl.gov) was established in 2002 through DOE Genomics:GTL funding of the Environmental Stress Pathway Project (ESPP). Using stress-response and metal reduction as target processes, and the environmentally important sulfate reducing bacteria (specifically *Desulfovibrio vulgaris* Hildenborough) as an organismal focus, ESPP is developing this system in two main thrusts: 1. Environmental characterization of biogeochemistry and microbial processes at metal/radionuclide contaminated sites, and 2. Laboratory and computational characterization of environmental stress pathways in microbes. Phase 2 of this project is increasing focus on molecular determinants of community activity, stability and ecology. This research is managed by three core teams: Applied Environmental Microbiology Core, Functional Genomics and Imaging Core, and Computational and Systems Biology Core.

DOE oversees 350 cleanup projects involving soil contaminated with metals/radionuclides. The life-cycle cost of these projects is at least $220 billion over 70 years, without breakthroughs. A thorough understanding of the biogeochemistry, especially stress responses in metal/radionuclide bacteria, enables prediction of natural attenuation and new strategies for remediation saving DOE billions in cleanup, risk assessment, and environmental stewardship. This application is representative of an array of environmental, ecological, and bioenergy stewardship challenges that rest on developing a detailed understanding of environmental microbial physiology, community interactions, population genetics and functions and ultimately evolution. The diversity of knowledge/technology necessary to accomplish these goals necessitates a team science approach, building a sophisticated experimental and computational infrastructure.

In the first phase of this project, we have succeeded in bringing a member of an important environmental class of microbe, the sulfate reducing bacteria (SRB), to nearly model organism status. In the last five years, we have been able to gain great insight into the stress response and metabolism of *Desulfovibrio vulgaris* Hildenborough (DvH). To do so we have developed an efficient genetic system; created a general purpose microbial systems biology pipeline; measured a wide range of genome-scale

* Presenting author

physiological responses to perturbations found in DOE contaminated environments; inferred regulatory systems and their evolution and conservation across metal-reducing SRB; and have just begun to probe DvH interactions with community members at DOE contaminated sites. The ESPP team has gelled as a mature working group of scientists able to balance the scale and coordination necessary for environmental systems biology with the special individual expertise to pursue detailed follow up projects. This synergy among labs and resources, and the economy of scale gained by building and utilizing a common experimental and computational infrastructure, enables attacking more complex and larger problems than any one or two labs could alone. This phase of the project has produced over 104 publications while team members have collectively made several hundred presentations of this work worldwide. The initial project has resulted in the development of a number of new technologies for functional genomics, data analysis, and physiological control of culture conditions. We have developed two community computational resources for comparative functional analysis of microbial genomes, MicrobesOnline (http://microbesonline.org) and RegTransBase (http://regtransbase.lbl.gov). We have 27 posters at this meeting outlining a number of the particular successes of this project as well as some of the work in preparation for phase two.

In phase two we will use this well-developed foundation to launch an expedition to discover the molecular mechanisms by which microbial community structure, function, and stability affects stress response and activity in the lab and in the field. With our current results in hand, we are poised to link field experiments strongly to controlled experiments with constructed consortia in the laboratory. We will continue to enrich our knowledge of the cellular networks of DvH and its relatives and expand our study to include its association with the methanogens with which it forms syntrophic associations. In stable co-cultures of *Desulfovibrio*-like organisms (Dvlo) and methanogens, we will discover the molecular basis of their metabolic coupling and other factors that lead to their more or less stable and active association leading to population growth, metal-reducing activity and resource utilization. We will construct more complex consortia to study the role of functional diversity, redundancy, and stress response in creating more or less stable microecologies via interactions at the community level. Finally, we will observe related microbes at DOE metal contaminated sites before, during, and after stimulation/stress to determine the differences and similarities to our laboratory experiments. There is little understanding of the molecular basis of microbial community formation, stability, resilience, specificity, activity and ecology, and we will be carrying out one of the first large-scale systems biology effort to explore these ubiquitous and important phenomena. To do so we are also expanding our characterization tools to include high throughput genetic techniques, new metabolomics methods for mixed cultures, new profiling technologies for understanding population and gene expression changes (functional metagenomics) over time in complex field studies, as well as the computational tools with which to manage, analyze and interrelated these data sets.

This new effort to discover the molecular determinants of community activity, stability, and ecology (MDCASE) is truly only possible at this scale of collaboration with nine institutions and over 75 full time scientists working together across the scales from single microbe to /in situ /environmental community function. This presents its own challenges in project management, communication, and balancing individual interests against project wide studies. Over the last five years we believe we have established an agile style that allows for investigators to pursue individual research while benefiting from and contributing to group projects and culture. In depth posters on the three cores and their accomplishments as well as the project management experience can be found nearby.

37

GTL

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

Terry C. Hazen[1,2*] (TCHazen@lbl.gov), Carl Abulencia,[1,4] Gary Anderson,[1,2] Sharon Borglin,[1,2] Eoin Brodie,[1,2] Steve van Dien,[8] Matthew Fields,[1,7] Julian Fortney,[1,2] Jil Geller,[1,2] E. Hendrickson,[5] Hoi-Ying Holman,[1,2] J. Leigh,[1,6] T. Lie,[1,6] Richard Phan,[1,2] Janet Jacobsen,[1,2] Dominique Joyner,[1,2] Romy Chakraborty,[1,2] **Martin Keller**,[1,3] **Aindrila Mukhopadhyay**,[1,2] Christopher Schadt,[1,3] **David Stahl**,[1,6] Sergey Stolyar,[1,6] Chris Walker,[1,6] **Judy D. Wall**,[1,5] Eleanor Wozei,[1,2] Zamin Yang,[1,3] Huei-che Yen,[1,5] Grant Zane,[1,5] and **Jizhong Zhou**[1,9]

[1]Virtual Institute of Microbial Stress and Survival (http://vimss.lbl.gov); [2]Lawrence Berkeley National Laboratory, Berkeley, California; [3]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [4]Diversa, Inc., San Diego, California; [5]University of Missouri, Columbia, Missouri; [6]University of Washington, Seattle,

* Presenting author

Washington; [7]Montana State University, Bozeman, Montana; [8]Genomatica, San Diego, California; and [9]University of Oklahoma, Norman, Oklahoma

Project Goals: See goals for abstract 36.

Field Studies

Environmental Characterizations. Clonal libraries were used to characterize changes in community structure along a contaminant plume (Oak Ridge, TN) in terms of phylogenetic, functional, and geochemical changes. Such studies are essential to understand how a microbial ecosystem responds to perturbations. Relationships between community diversity and ecosystem function were idiosyncratic, and these results suggested the population distributions depended on conditions under which the local landscape was investigated. Principal component analysis showed that nitrate, uranium, sulfide, sulfate, and COD were strongly associated with particular bacterial populations. Sequences closely related to nitrate-reducing bacteria were predominant during the initial phase of the remediation process, but sequences representative of sulfate-reducers (*Desulfovibrio* and *Desulfosporosinus spp.*) and metal-reducers (*Geobacter spp.*) were detected at higher levels as uranium levels declined. When engineering controls were compared to the community structure and composition via canonical ordinations, population distributions could be related with dissolved oxygen control and the presence of bio-stimulant. During the biostimulation, population distributions followed geochemical parameters, and these results indicated that bacteria exhibited distributions at the landscape scale in concordance with predictable geochemical factors. The data indicated that relationships between community structure and ecosystem function were idiosyncratic, but temporal and spatial concordance were eventually observed for the two bio-stimulated wells. In addition, evolution of the communities suggested that many populations were initially present and the changing geochemical conditions selected for multiple populations in a non-clonal fashion. Ultimately, sequences associated with sulfate-reducing populations predominated, and enrichment cultures are dominated by *Desulfovibrio* populations.

Previous research specifically points toward SRB as environmentally relevant experimental systems for the study of heavy metal and radionuclide reduction, and our recent data has detected *Desulfovibrio* sequences at the FRC and Hanford 100H. To effectively immobilize heavy metals and radionuclides, it is important to understand the cellular responses to adverse factors observed at contaminated subsurface environments, such as mixed contaminants and the changing ratios of electron donors and acceptors. In a recent study, we focused on responses to Cr(VI). At relatively low sulfide levels, *D. vulgaris* cells decoupled lactate oxidation from sulfate reduction for extended time periods even though all the Cr(VI) had been reduced. The cells could be protected by molecules that could complex Cr(III), and these results suggested that the reduction of Cr(VI) and/or reduction by-products were toxic to the cells. In addition, transcriptomic and physiological responses to Cr(VI) suggested that oxidizing stress was not a significant portion of the Cr induced stress, but was most likely a denatured protein response.

Technique Development for Environmental DNA and mRNA analysis. This year we did further optimization of the MDA approach to isolate and amplify DNA form samples with extremely low biomass. We used Hanford 100H samples to construct environmental libraries for sequencing and screening. We also evaluated three different methods in pure cultures to remove rRNA and tRNA from samples in order to screen mRNA expression that will eventually be applied to environmental samples. The first method utilizes biotin-modified oligos complementary to conserved regions in 16S and 23S rRNA and subtractive hybridization with streptavidin-coated magnetic beads. The second uses a commercially available exonulclease that specifically digests rRNAs bearing a 5' monophosphate group. The third method uses two rounds of reverse transcription, where rRNAs are first reverse transcribed with multiple universal primers for 16S and 23S RNAs, subsequently the RNA/DNA hybrids and cDNA are removed by sequential digestion with RNaseH and DNaseI, and the enriched mRNAs are then reverse transcribed using random primers. We evaluated these three methods by comparing disappearance of the 16S and 23S bands via electrophoresis, and their effect on mRNA quality and quantity by analysis of transcription levels of control (total RNA) vs. enriched mRNA as measured whole genome microarray. Enriched mRNAs from the first two methods generated more genes with altered transcript levels compared to untreated total RNA, with 19 genes (0.5%) for the exonuclease method and 74 genes (2%) for subtractive hybridization exhibiting significant differences ($P < 0.05$).

Genome Sequence. The genome sequence for *Desulfovibrio vulgaris* strain DePue has been closed and annotated in collaboration with other ESPP investigators. This genome appears highly conserved and syntenous compared with *D. vulgaris* Hildenborough except for dramatic differences in bacteriophage content. Additionally, strain DePue contains a large (~50 kb) unique region encoding for an exopollysaccahride production, modification and transport system. The genome of *Anaeromyxobacter* fw109 was recently completed and annotation

is underway. *Anaeromyxobacter* fw109-5 is a mesophilic, iron-reducing d-*Proteobacterium* that was recently isolated from subsurface sediments (approximately 15 m) at the ERSP-FRC in Oak Ridge, TN. Based upon SSU rRNA gene sequences, the closest cultivable relative is *Anaeromyxobacter dehalogenans* 2CP-C with 96.5% sequence identity. Approximately 105 putative proteins are estimated to contain heme-binding sites, with almost half being multi-heme proteins. This year we also submitted *D. vulgaris Miyazaki, D. salixigens*, and *D. desulfuricans* 27774 for sequencing at JGI. We are also preparing *D. termitidis, D. fructosivorans*, and two Hanford 100H *Desulfovibrio* sp. for sequencing at JGI.

Artificial communities. Transcriptional and mutational analyses of a coculture composed of *Desulfovibrio vulgaris* Hildenborough and *Methanococcus maripaludis* LL suggests *D. vulgaris* relies upon largely independent mechanisms of energy conservation for syntrophic growth and sulfate-respiration. In particular, the Coo hydrogenase, high-molecular weight cytochrome, [Fe] hydrogenase and [NiFe] hydrogenase play vital roles during syntrophic growth. Additionally, these studies provide a foundation for understanding the more complex communities described below. To better understand the ecology of the syntrophy, we have investigated the effect of genetic diversity on dual culture function by pairing *M. maripaludis* with different *Desulfovibrio* strains. These experiments showed that efficiency of coculture growth is correlated with numerical dominance of *Desulfovibrio* over *M. maripaludis* in co-cultures. Cocultures in which both species were equally predominant or where *M. maripaludis* dominated grew more slowly and achieved lower biomass. To further understand the ecological relevance of these phenotypic differences, we developed a protocol for competing *Desulfovibrio* strains against one another in syntrophic conditions and in monoculture with sulfate as the electron acceptor. To elucidate the potential genetic and phenotypic changes that may predominate in populations growing syntrophically for many generations, we founded 24 independent, but clonally identical populations and have allowed them to evolve for over 200 generations. Preliminary assessments of evolutionary changes indicate that populations evolved higher biomass by 100 generations of evolution, and have evolved increased stability. Currently, we are developing more complex assemblies to examine competitive and cooperative interactions, and factors contributing to community stability. We took advantage of the fully sequenced genomes of *Desulfovibrio vulgaris, Geobacter sulfurreducens* PCA, *Geobacter metallireducens* GS-15 and *Methanococcus maripaludis* to engineer various tri-cultures and develop tools to monitor community composition. Initial studies have shown that, tri cultures of *D. vulgaris*, *M. maripaludis* and *G. sulfurreducens* growing with fumarate and iron citrate and *G. metallireducens* growing with nitrate and iron citrate produced methane and some amount of hydrogen.

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. For the past five years, our ESPP 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. In the past five years we have produced biomass for 300 (150 in the last year) integrated experiments (oxygen, NaCl, NO_3, NO_2, heat shock, cold shock, pH, Cr, and mutants Fur, Zur, Per, and MP(-)) each with as much as 400 liters of mid-log phase cells (3×10^8 cells/ml). This year we have also focused on adapting these techniques to *Geobacter metallireducens* and have begun stress biomass production for comparison to *DvH* and *Shewanella*. 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. This has enabled us to link nitrogen utilization plate substrates to specific pathways for *D. vulgaris* as defined by the KEGG database. Characterization of the megaplasmid containing strain was confirmed by the plasmid sequence detailing the nitrogen fixation genes present on the plasmid. Crude mixed culture consisting of *D. vulgaris* and *G. metallireducens* showed PM profiles distinctly different from the profiles of the cultures run solo. We are in the process of modifying techniques to amplify signal of non-SRB organisms with low biomass for further phenotypic profiling.

Phenotypic Responses. We have generated a large set of phenotypic data that suggest analysis of the strain DePue genome sequence will provide important insights into the acquisition of metal-resistance absent in the closely related strain, *D. vulgaris* Hildenborough. An initial phenotypic characterization of a novel *Desulfovibrio* species isolated from the Hanford demonstration site has been completed and DNA is now being prepared for genome sequencing. We have completed extensive phenotypic comparisons of a large study set of *Desulfovibrio* species (14 different strains), as a prelude to continued comparative studies of fitness and evolution. Based upon

* Presenting author

transcriptomic data, a mutant was generated in a gene annotated as a sensory-box protein. Phenotypic analyses indicated that the mutant was deficient in stationary-phase survival. The lack of viability is caused by increases in the sulfide concentration. The data indicated the conserved hypothetical protein was important for sensing sulfide and directly or indirectly regulated a putative set of genes important for the cellular response to end-product accumulation.

Synchrotron FTIR Spectromicroscopy for Real-Time Stress Analysis. Synchrotron FTIR Spectromicroscopy approaches will be discussed in another presentation at this meeting.

38

GTL

Applications of Systems Biology Approaches to Understanding Artificial Microbial Consortia and Environmental Communities in the VIMSS Applied Environmental Microbiology Core

Christopher Schadt[1,4]* (schadtcw@ornl.gov), Zamin Yang,[1,4] Amudhan Venkateswaran,[1,4] Meghan Drake,[1] Susan Carroll,[1,4] Dawn Klingeman,[1,4] Mircea Podar,[1,4] Thomas Phelps,[1,4] Steven Brown,[1,4] Anthony Palumbo,[1,4] Sergey Stolyar,[2,4] Christopher Walker,[2,4] **David Stahl**,[2,4] **Terry C. Hazen**,[3,4] and **Martin Keller**[1,4]

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [2]Department of Civil and Environmental Engineering, University of Washington, Seattle, Washington; [3]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; and [4]VIMSS (Virtual Institute of Microbial Stress and Survival) http://vimss.lbl.gov/

Project Goals: See goals for abstract 36.

Cultivation of single species has been at the central core of experimental microbiology for more than a century but offers only a glimpse into the biology of microorganisms in nature. Communities, not individual species, control the process rates that drive key biogeochemical cycles, including the transformation of environmental pollutants of concern to DOE. Thus detailed studies of model consortia and communities that mediate such processes that will allow for experimental manipulation and in-depth analysis of the fundamental biology underlying such systems are essential for advancing DOE objectives. As part of the GTL Environmental Stress Pathway Project (ESPP) team, we are pursuing two projects to advances these objectives.

Methods development for environmental mRNA analysis. Current technologies applied to environmental samples for RNA transcriptional profiling include RT-PCR and functional gene microarrays. While, tremendous progress has been made in understanding microbial communities due to emergence of these technologies, they bear significant limitations that prevent their application in a high throughput manner to *de novo* communities. We are developing methods for directly sequencing cDNA from environment samples utilizing new high throughput (HT) sequence analysis technologies. Since 80% or more of total RNA from bacteria is represented by the rRNA pool, it is crucial to first remove those components as thoroughly as possible without adversely impacting mRNA quality, quantity and composition, prior to HT sequence based screening. We have compared three different methods to remove rRNAs and enrich mRNAs of ***D. vulgaris*** Hildenborough (DvH) samples. The first method utilizes biotin-modified oligos complementary to conserved regions in 16S and 23S rRNA and specific removal by binding to streptavidin-coated magnetic beads. The second uses a commercially available exonuclease that specifically hydrolyzes rRNAs bearing a 5' monophosphate group. The third method uses two rounds of reverse transcription, where rRNAs are first reverse transcribed with multiple universal primers for 16S and 23S rRNAs and subsequently the RNA/DNA hybrids and cDNA are removed by sequential digestion with RNaseH and DNaseI. We have evaluated these three methods alone and in combination using microarray-based analysis of transcription levels. All three methods are able to significantly enrich mRNA from rRNA without introducing significant biases. Microarray analysis revealed significant differences in measured mRNA levels in only 0.5% to 5% of genes across the genome as compared to controls. Comparisons of microarray results with HT sequencing using the Solexa platform are currently ongoing. After further validation, application of these methods could be performed on environmental systems from the Hanford and/or Oak Ridge contaminated sites as part of the VIMSS/ESPP applied environmental core studies, as a compliment the DNA based metagenomic analyses already underway at these sites.

Developing manipulatable, laboratory based, higher order microbial consortia. A practical understanding of how community structure leads to process rates and stability is central to DOE objectives in bioremediation and process

* Presenting author

control. Although there are numerous theories relating to stability in macroecology, their relevance to microbial communities is mostly untested. To further these studies we are in the initial stages of assembling model microbial consortia in the laboratory that will allow us to study and manipulate community interactions in a controlled manner and test the stress responses of the assemblages. The model organisms now used by the VIMSS/ESPP team will serve for constructing initial consortia, encompassing sulfate-reducers (*DvH*), iron and uranium reducers (*G. metallireducens*) and methanogens (*M. maripaludis*). The genomes of all these strains have been sequenced, gene expression microarrays are available within the group, and the individual organisms can be genetically manipulated, that will allow unprecedented toolsets to be applied to these controlled communities that would not be possible in natural systems. Additional methanogens, metal-reducing bacteria from DOE contaminated sites (e.g. *Geobacter*, *Anaeromyxobacter*, and *Desulfovibrio* sp.) and heterotrophic *clostridia* that can provide end products of cellulose fermentation (ethanol, acetate and lactate) to the other community members are also to be added to consortia designs. Three member consortia combinations including *C. acetobutylicum*, DvH and *G. lovelii,* as well as, additional consortia with DvH, *G. metallireducens*, and *M. maripaludis* have been tested thus far in the past few months. Methods for tracking population dynamics of the consortia members such as qPCR and FISH have also been developed and have shown relatively stable assemblages of these species can be achieved. Higher order consortia studies incorporating 4 and more community members are ongoing.

Together these studies will enable us to do in depth studies of stress mechanisms within environmental and model consortia systems, and understand how the detailed mechanisms outlined using pure culture laboratory systems within the Functional Genomics Core of the VIMSS/ESPP project, translate into the relationships and activities observed in more complex constructed consortia as well as ultimately into environmental microbial communities.

39

GTL

Growth Rate and Productivity of a Microbial Mutualism Depends on the *Desulfovibrio* Genotype

Kristina L. Hillesland* (hilleskl@u.washington.edu) and **David A. Stahl**

Department of Civil and Environmental Engineering, University of Washington, Seattle, Washington and VIMSS (Virtual Institute of Microbial Stress and Survival) http://vimss.lbl.gov/

Project Goals: See goals for abstract 36.

In a syntrophic mutualism, two species work together to gain energy from thermodynamically unfavorable reactions. This energy is partitioned between the species, but it is unclear how this occurs and if it is a parameter that can vary depending on the phenotype of each mutualist. In a genetically variable population, genotypes that are capable of acquiring a greater share of the energy from the syntrophic interaction may have a competitive advantage. If such unequal distribution has a negative effect on the overall efficiency of the mutualism, then natural selection could cause instability in these associations. Thus, it is important to understand the relationships between the overall efficiency of syntrophic growth, the allocation of resources among mutualists, and whether these relationships are genetically based. To explore these possibilities, we paired *Methanococcus maripaludis* with several species of the sulfate reducer *Desulfovibrio* and cultured them in the absence of an electron acceptor for *Desulfovibrio*, with lactate as electron donor. In these conditions, *Desulfovibrio* ferments lactate and produces hydrogen that *M. maripaludis* consumes for growth, thereby maintaining thermodynamically favorable conditions for lactate fermentation. We found that growth rate and biomass produced during syntrophic growth varied considerably depending on the genotype of *Desulfovibrio*. To elucidate the relationship between co-culture growth efficiency and composition, we used a study set of four co-cultures that varied in growth rate and biomass yield to track the population dynamics of each species. We estimated the abundance of each species at several time intervals from the relative concentrations of 16S rRNA. Co-cultures containing *D. desulfuricans* 27774 or *D. vulgaris* Hildenborough grew more quickly and achieved a higher OD_{600nm} than co-cultures containing either *D.* G20 or *D. vulgaris* oxamicus. In the faster growing co-cultures, *Desulfovibrio* dominated over *Methanococcus* at least 2-fold, but the relative numbers in the slower growing co-cultures were roughly equal or

* Presenting author

favored *Methanococcus*. These results are consistent with results of our flux-balance model of a two species association, which showed that *Desulfovibrio* should dominate 2-2.5-fold over *Methanococcus* in syntrophies that are growing optimally. Thus, unequal allocation of energy towards *Desulfovibrio* may be explained by fluxes through the stoichiometric metabolic network underlying the mutualistic relationship, and may even benefit *Methanoccous* by enabling faster growth.

40

GTL

Biodiversity and Spatial Concordance of an in Situ System for Uranium Bioreduction

C. Hwang,[1,7] W.-M. Wu,[2] T.J. Gentry,[3] J. Carley,[4] S.L. Carroll,[4] D.B. Watson,[4] P.M. Jardine,[4] **J. Zhou**,[5,8] C.S. Criddle,[2] and **M.W. Fields**[6,7,8]* (matthew.fields@erc.montana.edu)

[1]Department of Microbiology, Miami University, Oxford, Ohio; [2]Department of Civil and Environmental Engineering, Stanford University, Stanford, California; [3]Department of Crop and Soil Sciences, Texas A&M University, College Station, Texas; [4]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [5]Institute for Environmental Genomics, University of Oklahoma, Norman, Oklahoma; [6]Department of Microbiology and [7]Center for Biofilm Engineering, Montana State University, Bozeman, Montana; and [8]Virtual Institute of Microbial Stress and Survival http://vimss.lbl.gov/

Project Goals: See goals for abstract 36.

The elucidation of how populations of interest interact in a given community and how the community responds to stress and perturbations can help infer the interplay between stress pathways and gene networks that help optimize bacterial biochemistry. A goal of VIMSS is to characterize the responses of bacterial communities at multiple levels of resolution in order to understand biochemical capacity at DOE waste sites. The current work uses a series of re-circulating wells that create a subsurface bioreactor to stimulate microbial growth for *in situ* U(VI) immobilization (Wu et al. ESandT 41:5716-5723) and the bacterial communities were tracked over time and space in response to nutrient influx. Bacterial community dynamics were investigated in a series of re-circulating wells that created a subsurface "bio-reduction zone" to stimulate bacterial growth with ethanol for in situ bioremediation of U(VI) at the Field Research Center of the U.S. Department of Energy, Oak Ridge, TN. Different experiments were conducted to alter the subsurface environment to better understand strategies that would improve the remediation process. Within this framework, the interrelationships between the biogeochemistry were studied in order to characterize the community and ecosystem ecology with respect to microbiology of an engineered system. Bacterial community composition and structure of groundwater samples were analyzed via clone libraries of partial SSU rRNA genes. UniFrac analyses showed that the bacterial community in each of the wells developed changes during the bioremediation process, and the changes could be attributed to the variations along temporal and spatial scales. Relationships between community diversity and ecosystem function were idiosyncratic, and these results suggested the population distributions depended on the particular conditions under which the local landscape was investigated. Principal component analysis showed that nitrate, uranium, sulfide, sulfate, and COD were strongly associated with particular bacterial populations. Sequences closely related to nitrate-reducing bacteria were predominant during the initial phase of the remediation process, but sequences representative of sulfate-reducers (*Desulfovibrio* and *Desulfosporosinus spp.*) and metal-reducers (*Geobacter spp.*) were detected at higher levels as uranium levels declined. Ultimately, sequences associated with sulfate-reducing populations predominated. Uranium levels declined below EPA drinking water standards, and community composition and structure were similar in both treatment wells after approximately 1.5 y despite going through different transitions. In addition, when engineering controls were compared to the community structure and composition via canonical ordinations, population distributions could be related with dissolved oxygen control and the presence of bio-stimulant. During the bio-stimulation, population distributions followed geochemical parameters, and these results indicated that bacteria exhibited distributions at the landscape scale in concordance with predictable geochemical factors. The data indicated that relationships between community structure and ecosystem function were idiosyncratic, but temporal and spatial concordance were eventually observed for the two bio-stimulated wells. The strong associations between particular environmental variables and certain population distributions will provide insights into establishing practical and successful remediation strategies in radionuclide-contaminated environments with respect to engineering controls and ecosystem function.

* Presenting author

41

GTL

Studying Rules Governing Microbial Communities

Sergey Stolyar* (sstolyar@u.washington.edu) and **David A. Stahl**

Department of Civil and Environmental Engineering, University of Washington, Seattle, Washington and Virtual Institute for Microbial Stress and Survival (http://vimss.lbl.gov)

Project Goals: See goals for abstract 36.

Understanding the behavior of biological communities presents tremendous challenges because of the complex network of diverse interactions among species. A class of communities of particular interest from ecological, geological, and engineering perspectives is represented by microbial communities that thrive in oxygen-free (anoxic) environments. These communities are vital components in numerous environments including freshwater sediments, the subsurface, guts of insects and animals, wastewater treatment plants and heavy-metal contaminated environments of concern to DOE. They therefore play a significant role in global cycling of carbon and other biogenic elements. Our prior research established the feasibility of working with a simple two-tier food web composed of two species, each species occupying a distinct trophic position: *Desulfovibrio vulgaris* syntrophically coupled to the hydrogen consuming *Methanococcus maripaludis*. As part of phase two of the Environmental Stress Pathway Project (ESPP) we are now developing more complex assemblies to examine competitive and cooperative interactions, and factors contributing to community stability. We are using organisms having fully sequenced genomes (*Desulfovibrio vulgaris, Geobacter sulfurreducens* PCA, *Geobacter metallireducens* GS-15 and *Methanococcus maripaludis*) to construct different tricultures and develop tools to monitor community composition. *G. sulfurreducens* and *G. metallireducens* consume acetate and use alternative electron acceptors, including nitrate, fumarate and iron. The ecological relevance of this electron acceptor versatility in constructed microbial food webs is being examined by measuring growth kinetics and yields in relationship to the modynamic predictions. Initial studies have shown that lactate-grown tricultures of *D. vulgaris* and *M. maripaludis* combined with either *G. sulfurreducens* (plus fumarate or iron citrate) or *G. metallireducens* (plus nitrate or iron citrate) evolved both methane and hydrogen at different stages of growth in batch culture. Tri-cultures with *G. sulfurreducens* growing with fumarate and tri culture with *G. metallireducens* growing with nitrate demonstrated comparable growth rates and biomass yields despite differences in predicted total free energy: -671 kJ for fumarate and -1183 kJ for nitrate.

42

GTL

Energy Conservation in a Biogeochemically Significant Microbial Mutualism

Christopher B. Walker,[1,8,*] Zhili He,[2,3,8] Zamin K. Yang,[2,8] Janet Jacobsen,[4,8] Joseph A. Ringbauer, Jr.,[5,8] Qiang He,[2,8] **Jizhong Zhou**,[2,3,8] Gerrit Voordouw,[6] **Judy D. Wall**,[5,8] **Adam P. Arkin**[4,7,8] (aparkin@lbl.gov), **Terry C. Hazen**,[4,8] Sergey Stolyar,[1,8] and **David A. Stahl**[1,8]

[1]Department of Civil and Environmental Engineering, University of Washington, Seattle, Washington; [2]Environmenal Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [3]Institute For Environmental Genomics and Department of Botany and Microbiology, University of Oklahoma, Norman, Oklahoma; [4]Lawrence Berkeley National Laboratory, Berkeley, California; [5]University of Missouri, Columbia, Missouri; [6]University of Calgary, Calgary, Alberta, Canada; [7]Department of Bioengineering, University of California, Berkeley, California; and [8]Virtual Institute for Microbial Stress and Survival (http://vimss.lbl.gov)

Project Goals: See goals for abstract 36.

Complex interactions govern microbial communities in both pristine and contaminated environments. Unfortunately, a limited understanding exists regarding the interactions sustaining these communities. Without a deeper knowledge of the molecular basis driving the community structure and function, bioremediation of contaminated sites suffers from inefficient or ineffective design strategies. The VIMSS/ESPP2 project aims at resolving the molecular basis for microbial organisms and communities found at heavy-metal contaminated environments, such as the Hanford 100H site. Sulfate-reducing microbes (SRM) commonly compose significant fractions of the microbial community in contaminated anaerobic sites. The VIMSS/ESPP2 project extensively examined a representative SRM, *Desulfovibrio vulgaris* Hildenborough, building a detailed understanding of stress response mechanisms through intensive monoculture study. However, *D. vulgaris* often populates environments deficient in sulfate, relying upon syntrophic associations with

* Presenting author

hydrogenotrophic methanogens for continued growth. Investigation of an archetypical community composed of *D. vulgaris* Hildenborough and a representative methanogen, *Methanococcus maripaludis* LL, serves as a basis for understanding the physiological differences between growth modalities. Using transcriptional analysis, we demonstrate that continuously grown cultures of *D. vulgaris* Hildenborough, up-regulate a broad suite of electron transfer enzymes during syntrophic growth. Mutational analyses indicate keys roles for four enzymes (Coo, Hmc, Hyd and Hyn) not essential for sulfate-respiration. Specifically, these results provide a developed molecular basis for understanding this "community of two" while also serving as the foundation for future VIMSS/ESPP2 community analysis. More generally, these results suggest syntrophic growth and sulfate-respiration rely upon largely independent energy generation pathways.

43

GTL

ESPP Functional Genomics and Imaging Core: Cell Wide Analysis of Metal-Reducing Bacteria

Aindrila Mukhopadhyay,[1,6]* Edward Baidoo,[1,6] **Kelly Bender**[5,6] (bender@micro.siu.edu), Peter Benke,[1,6] Swapnil Chhabra,[1,6] Gabriela Chirica,[2,6] Elliot Drury,[3,6] Dwayne Elias,[3,6] Veronica Fok,[1,6] Sara Gaucher,[2,6] Masood Hadi,[2,6] Zhili He,[4,6] Chris Hemme,[4,6] **Jay Keasling**[1,6] (keasling@.berkeley.edu), Kimberly Keller,[3,6] Yooli Light,[2,6] Eric Luning,[1,6] Robert Meagher,[2,6] Francesco Pingitore,[1,6] Alyssa Redding,[1,6] Jarrod Robertson,[3,6] Rajat Sapra,[2,6] Matthew Shirley,[3,6] **Anup Singh**,[2,6] (aksingh@sandia.gov), Yinjie Tang,[1,6] **Judy Wall**,[3,6] (wallj@missouri.edu), Huei-Che Yen,[3,6] Grant Zane,[3,6] Aifen Zhou,[4,6] and **Jizhong Zhou**[4,6] (jzhou@rccc.ou.edu)

[1]Lawrence Berkeley National Laboratory, Berkeley, California; [2]Sandia National Laboratories, Livermore, California; [3]University of Missouri, Columbia, Missouri; [4]University of Oklahoma, Norman, Oklahoma; [5]Southern Illinois University; and [6]Virtual Institute of Microbial Stress and Survival http://vimss.lbl.gov

Project Goals: See goals for abstract 36.

A fundamental goal of the Environmental Stress Pathway Project (ESPP) of the Virtual Institute of Stress and survival (VIMSS) is a rigorous understanding of *Desulfovibrio vulgaris* Hildenborough physiology and its ability to survive in its environment. Such knowledge will be critical in discerning the biogeochemistry at metal contaminated sites, for bioremediation and natural attenuation for toxic metals. The Functional Genomics and Imaging Core (FGIC) focuses on the measurement of these responses at a cell wide level using systems biology approaches.

Progress in the last one year built upon our optimized pipeline for generating biomass for various functional genomics studies and utilized improved genetic methods. Numerous additional transcriptomics data sets were added to our compendium of stress response studies. These included peroxide stress, low oxygen stress, high and low pH stress and alteration of growth conditions (e.g presence of methionine, alternate electron donors etc). To understand how genotype and environment interact to determine the phenotype and fitness of an organism, a long-term evolution experiment was also conducted to examine the dynamics and adaptation of *D. vulgaris* under extended salt exposure. For many of these stresses iTRAQ based quantitative proteomics and CE-MS based metabolite studies were also conducted. Improved genetic methods were employed to create several critical knock out mutants (e.g. *echA*, *qmoABC and tatA*) and several were characterized via growth and transcriptomics studies. Progress was also made in extending transcriptomics analysis to examine alternate *D. vulgaris* physiological states such as in biofilms and growth in syntrophic co-culture with *Methanococcus maripaludis*. Methods to conduct iTRAQ proteomics and stable isotopomer (^{13}C) based metabolic flux analysis were also developed for studying co-cultures. A novel FTICR-MS based method for a comparative $^{12}C/^{13}C$ based metabolite analysis is being developed and will enable a direct comparison of control cultures to experimental samples. Additionally we continued to collect cell wide data in *Shewanella oneidensis* and *Geobacter metallireducens* for comparative studies. Great progress was made in improving extraction and high throughput of metabolite studies. Metabolite extraction and CE-MS detection for several hundred metabolites can now be conducted for these non-model organisms using high resolution separation and high resolution mass spectrometric methods.

Continued studies to map cell wide responses have also emphasized the importance of changes that require orthologous measurements. With this in mind a novel protocol to monitor protein-protein interactions and redox state of the proteins has been developed. In an effort to supplement model development and elucidate intricacies of stress response cascades, comprehensive methods for identifying alternative regulatory mechanisms such as small non-coding RNAs are also underway. To optimize the use of the large amounts of data being

* Presenting author

collected, several data mining efforts were initiated. For example, iTRAQ data sets from the multiple stress response studies were mined for potential post translational modifications and confirmation of hypothetical proteins while ^{13}C flux data were used to confirm gene annotation and assess missing steps in metabolic pathways. Work in underway in collaboration with the computational core to set up searchable databases of our proteomics and metabolite data also.

44

GTL

VIMSS ESPP: Deciphering the Roles of Two-Component Systems in *Desulfovibrio vulgaris* Hildenborough

Aindrila Mukhopadhyay[1,2*] (AMukhopadhyay@lbl.gov), Dominique Joyner,[1,3] Eric Luning,[1,2] Kimberly Keller,[1,4] Jarrod Robertson,[1,4] Grant Zane,[1,4] Janet Jacobsen,[1,2] Morgan Price,[1,2] Swapnil Chabbra,[1,2] **Terry C. Hazen**,[1,3] **Adam P. Arkin**,[1,2] **Judy Wall**,[1,4] and **Jay Keasling**[1,2]

[1]Virtual Institute of Microbial Stress and Survival; http://vimss.lbl.gov; [2]Physical Biosciences Division and [3]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; and [4]Biochemistry and Molecular Microbiology and Immunology Departments, University of Missouri, Columbia, Missouri

Project Goals: See goals for abstract 36.

One of the primary goals of the Environmental Stress Pathway Project (ESPP) is to map the response of the anaerobic sulfate reducing soil bacterium, *Desulfovibrio vulgaris* Hildenborough to its environment. Two component systems, comprised of Histidine Kinase and Response regulator proteins, present the primary and ubiquitous mechanism in bacteria for initiating cellular response towards a wide variety of environmental conditions. In *D. vulgaris* Hildenborough, more than 70 such systems have been predicted, but remain mostly uncharacterized. The ability of *D. vulgaris* to survive in its environment is no doubt linked with the activity of genes modulated by these two component signal transduction systems. To map the two component systems to the genes they modulate, the availability of deletion mutants provides an important tool. Here we present an overview of the predicted histidine kinases in *D. vulgaris* and describe a strategy to create library of histidine kinase knock out mutants in *D. vulgaris*. We use the OmniLog® workflow to conduct a wide phenotypic characterization of the knock out mutants generated. To illustrate our strategy we present results from our study of the histidine kinase in the predicted *kdp* operon of *D. vulgaris*. The high-affinity potassium uptake Kdp complex is well characterized in other bacteria where it facilitates K^+ uptake in low K^+ or high Na^+ conditions. Typically, the activity of the Kdp system is modulated by the KdpD/E two-component signal transduction system, where KdpD is the sensor histidine kinase and KdpE is the response regulator. The *D. vulgaris kdp* operon contains a gene with predicted response regulator function and two separate genes annotated for the sensor kinase function (*kdpD* and DVU3335). Interestingly, only one of these two, DVU3335, contains a conserved histidine kinase domain which is absent the *D. vulgaris kdpD* candidate. However, DVU3335 does not encode the well-conserved motifs associated with KdpD. We created a knock out mutant in the DVU3335 gene. The DVU3335 knock out strain showed a growth deficiency in low K^+ conditions and when exposed to low K^+ conditions was unable to upregulate genes in the *kdp* operon. Phenotypic microarrays were used to obtain a broader comparison of the mutant and wild type strains. Our results show that the major differences between the wild type and the mutant are in response to salt stress and support the role of DVU3335 in modulating K^+ uptake during low K^+ and high Na^+ conditions.

45

GTL

Observing Polyglucose Metabolism and Transient Oxygen Stress in Obligate Anaerobes in Vivo

Hoi-Ying N. Holman[1,2*] (hyholman@lbl.gov), Eleanor Wozei,[1] Zhang Lin,[1,3] Luis Comolli,[1] Kenneth H. Downing,[1] **Matthew Fields**,[2,4] and **Terry C. Hazen**[1,2]

[1]Lawrence Berkeley National Laboratory, Berkeley, California; [2]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov; [3]Fujian Institute of Research on Structure of Matter, Chinese Academy of Science, China; and [4]Montana State University, Bozeman, Montana

Project Goals: See goals for abstract 36.

Our purpose is to provide the VIMSS/ESPP(1) project with molecular information which will improve the existing understanding of microbial adaptive response strategies, and also will enable a better management and modeling of microbes in subsurface environments for cleanup of hazardous waste sites. Specifically, we studied

* Presenting author

at a chemical level how the obligate anaerobic *Desulfovirbio vulgaris* could survive in a variety of environments which periodically become aerobic.

Aerobic respiration of intracellular polyglucose reserves is postulated to play a central role in oxygen adaptive response in obligatory anaerobes like *Desulfovibrios*, but it has been difficult to probe this event at chemical scale *in vivo*. Here we presented a non-invasive synchrotron infrared (SIR) spectromicroscopy approach to reveal time-dependent composition and structure changes at a lateral scale of several individual *D. vulgaris*. The advantage of infrared spectroscopy is that it is non-invasive, and it uses vibration movements of atoms and chemical bonds within functional groups of biomolecules as an intrinsic contrasting mechanism; thus it allows one to immediately detect composition and structure changes within cells. The advantage of using a synchrotron light source is that its high brightness allows us to detect signals ~1000 times weaker then the conventional infrared spectroscopy allows us to detect.

Comparative analysis of SIR spectra of the same individual *D. vulgaris* exposed to air-level oxygen at different time points reveals chronological information regarding the level of oxidative stress and the extent of cellular injury and repair. These results, together with high-resolution microscopy images, mark a critical step toward the use of SIR spectromicroscopy as an uninterrupted microprobe at a chemical scale level of physiological events in microbiology applications.

46

GTL

Phenotypic Characterization of Microorganisms by Barcoded Transposon Mutagenesis

Adam Deutschbauer,[1,4] Julia Oh,[2] Morgan Price,[1,4] Paramvir Dehal[1,4]* (PSDehal@lbl.gov), Dan Bruno,[2] Marlene Henriquez,[1,4] Romy Chakraborty,[1,4] **Terry C. Hazen**,[1,4] Corey Nislow,[3] Guri Giaever,[3] Ronald W. Davis,[2] and **Adam P. Arkin**[1,4]

[1]Lawrence Berkeley National Laboratory, Berkeley, California; [2]Stanford Genome Technology Center, Stanford, California; [3]University of Toronto, Toronto, Ontario, Canada; and [4]VIMSS (Virtual Institute of Microbial Stress and Survival) http://vimss.lbl.gov/

Project Goals: See goals for abstract 36.

The Environmental Stress Pathway Project (ESPP) aims to elucidate the molecular mechanisms by which microbial communities affect stress response and activity in the field and in the laboratory using sulfate reducing bacteria (SRB) as a model. To accomplish our goal of linking field observations to those in the laboratory, a systems-level understanding of SRB genome function is necessary. To meet this challenge, we are developing a mutagenesis and phenotyping strategy that is comprehensive across the genome and applicable to any microorganism amenable to transposon mutagenesis. Here we describe the application of our strategy to *Shewanella oneidensis* MR1 and the SRB *Desulfovibrio desulfuricans* G20. We have cloned and sequence-verified ~3000 barcode modules into a Gateway entry vector. Each module is a 175 base pair element containing two unique 20 base pair sequences, the UPTAG and DOWNTAG, flanked by common PCR priming sites. Each module can then be rapidly transferred *in vitro* to any DNA element, such as a transposon, that is made Gateway compatible. Transposon mutants marked by the modules will be sequenced to determine which of the ~3000 barcode modules was used and which gene was disrupted. Transposon mutants can be rapidly re-arrayed into a single pool containing ~3000 uniquely tagged, sequence-verified mutant strains. By sequencing saturating numbers of transposon mutants, we can identify and assay mutants in most nonessential genes in a given genome. The fitness of each mutant in the pool will be monitored in parallel by the hybridization of the barcodes to an Affymetrix microarray containing the barcode complements in a system identical to that used for the yeast deletion collection. Compared to other approaches for the parallel analysis of transposon mutants such as signature tagged mutagenesis, genetic footprinting, and transposon site hybridization, our approach offers much higher throughput, a single microarray design is universal for any organism, single mutational events are assayed, and mutant strains are archived for verification, distribution, and the systematic genetic interrogation of individual pathways.

The successful completion of this project will enable the quantitative phenotypic analysis of thousands of mutants across a wide range of conditions. These data will be used to assign gene function on a global scale, aid in the identification of missing metabolic enzymes, and provide insight into the functional connectivity of different pathways. Our results in *Desulfovibrio desulfuricans* G20 will be extrapolated *via* comparative genomic analysis to other sequenced SRBs including *Desulfovibrio vulgaris* Hildenborough. Consequently, our findings will aid in the interpretation of both laboratory and field data collected by the ESPP.

* Presenting author

47

GTL

The Development of a Markerless Deletion System in *Desulfovibrio vulgaris* Hildenborough

Kimberly L. Keller[1,3]* (kellerkl@missouri.edu), **Kelly S. Bender**,[2,3] and **Judy D. Wall**[1,3] (WallJ@missouri.edu)

[1]Biochemistry Department, University of Missouri, Columbia, Missouri; [2]Microbiology Department, Southern Illinois University, Carbondale, Illinois; and [3]VIMSS (Virtual Institute of Microbial Stress and Survival) http://vimss.lbl.gov/

Project Goals: See goals for abstract 36.

To fully explore microbial community dynamics, stability of the composition with time and changing nutritional and environmental factors must be explored. In order to confirm sources and sinks of metabolites, both during degradation and biosynthesis, it would be most useful to create pivotal deletions in various members of the community. We are pursuing genetic tools that can possibly be applied to strains with limited genetic accessibility. These tools are being developed in *Desulfovibrio* and include an inframe deletion procedure and plasmid modification in extracts to facilitate genetic exchange processes.

To properly study metabolic pathways, it is necessary to delete several genes that may have compensatory activities. Our model system, the sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough has seen enormous progress in genetic manipulation; however, the current deletion method of marker exchange mutagenesis does not allow for easy selection of multiple sequential gene deletions because of the low number of selectable markers available in *D. vulgaris*. To broaden the repertoire of genetic tools available for manipulation in *D. vulgaris*, an in-frame markerless deletion system is being developed based on the *upp*-encoded uracil phosphoribosyltransferase as an element for a counterselection strategy. In wild-type *D. vulgaris,* growth is inhibited by the toxic pyrimidine analog 5-fluorouracil (5-FU), whereas a mutant bearing a deletion of the *upp* gene is resistant to 5-FU. The introduction of a plasmid containing the wild-type *upp* gene expressed constitutively from the *aph*(5')–III promoter (the promoter for the kanamycin resistance gene in Tn*5*) into the *upp* deletion strain restored sensitivity to 5-FU. This observation is the basis for the establishment of a two-step integration and excision strategy for the deletion of genes of interest. Since this deletion does not contain an antibiotic cassette, multiple gene deletions can be generated in a single strain using this method.

This in-frame markerless deletion method is currently being evaluated through the construction of a deletion of the putative formate dehydrogenase alpha- and beta- subunits, DVU0587 and DVU0588. In addition, Gateway Technology methods are being developed that would expedite the process of generating the required deletion vectors by the construction of a destination vector containing the constitutively expressed wild-type *upp* gene. This new method is being utilized to generate a deletion for the R-subunit (DVU1703) of a type I restriction-modification system.

48

GTL

Expression Profiling and Gene Association of Hypothetical and Conserved Hypothetical Genes in *Desulfovibrio vulgaris* Leads to Functional Annotation

Dwayne A. Elias,[1,5]* Elliot C. Drury,[1,5] Alyssa M. Redding,[2,5] **Aindrila Mukhopadyay**,[2,5] Huei-Che B. Yen,[1,5] Katherine H. Huang,[3,5] **Terry C. Hazen**,[2,5] **Adam P. Arkin**[2,4,5] (aparkin@lbl.gov), and **Judy D. Wall**[1,5]

[1]Department of Biochemistry, University of Missouri, Columbia, Missouri; [2]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; [3]Massachusetts Institute of Technology, Cambridge, Massachusetts; [4]Department of Bioengineering, University of California, Berkeley, California; and [5]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

Project Goals: See goals for abstract 36.

The annotation of predicted open reading frames within sequenced bacterial genomes has derived from physiological experimentation of pure bacterial cultures. This has resulted in a large residual fraction of unannotated genes. As studies move to more and more environmentally relevant microbial communities, these hypothetical and conserved hypothetical genes may well prove to encode critical functions. We have to date observed numerous hypothetical and conserved genes that respond to a single stress or condition, including in a sulfate-reducing bacteria (SRB)/methanogen co-culture. It is expected that such information will be useful in

* Presenting author

the shorter-term ESPP2 project goal of sequencing and annotating additional SRB from DOE sites, as well as the longer-term project goals of understanding the role of SRB in multi-organism and syntrophic cultures, including targeting genes for deletion that may be important in such conditions.

Hypothetical and conserved hypothetical genes consistently make up 30% or more of sequenced bacterial genomes, with few reports confirming their expression at either the rRNA or protein level. This is the case for the SRB *Desulfovibrio vulgaris* Hildenborough where 348 of the total 3534 genes are currently annotated as conserved or conserved hypothetical (9.7%) with an additional 889 hypothetical genes (25.0%). Given this large complement of the genome, it is plausible that some of these genes serve significant cellular functions that may well range from regulation to presently unknown steps in carbon or electron flux. The goals of this study were to determine which of these genes actually encodes a protein and then to infer a more functionally based annotation. In order to accomplish this, we have compiled and explored the microarray and MS-based iTRAC proteomic expression profiles for the expected 1237 hypothetical and conserved hypothetical proteins in *D. vulgaris* from the ten environmental stresses for which data have been collected in the VIMSS/ESPP project. Categories for examining the expression data included the predicted operon arrangements (monocistronic versus polycistronic operons), the basal transcription levels, and any differential expression in stressed cells. Overall, we are presently able to confirm the expression of 1219 genes at the mRNA level and 265 genes at both the mRNA and protein level, while there was no evidence for either mRNA or protein detectable for 17 genes. While the number of microarray studies outweighs proteomics, the abundance values indicating differential expression at the protein level were consistent with the microarray results when data were available. The monocistronic genes were then reannotated based on several factors including their stress response profile and COG information when available, while polycistronic genes were reannotated using the expression profiles, COG information and association with better annotated genes within the operon. Additionally, analyses of deletion mutants of *fur* and *perR,* encoding global regulatory systems revealed a number of genes apparently regulated directly or indirectly by Fur and/or PerR.

The validity of such inferred assignments can only be ascertained by interruption or deletion of the gene with further analysis. To this end, two targeted deletion mutants were constructed and used as test cases to determine the accuracy of the revised functional annotations. In each case the phenotype obtained was in agreement with the expanded annotation. We are also in the process of testing several mutants that have been isolated from a random transposon library. Through these efforts, a more precise inter-organism comparison will be possible, thereby yielding better clues as to the function of conserved or hypothetical genes in other organisms. Finally, some of the physiological and metabolic aspects of *D. vulgaris* that are currently not understood may be solved thus aiding bioremediative efforts by understanding the mechanisms by which this bacterium survives the stresses and community effects likely experienced at DOE contaminated sites.

49

GTL

Identification of a Small Non-Coding RNA in *Desulfovibrio vulgaris*

Kelly S. Bender[2,3]* (bender@micro.siu.edu) and **Judy D. Wall**[1,3]

[1]Biochemistry Department, University of Missouri, Columbia, Missouri; [2]Microbiology Department, Southern Illinois University, Carbondale, Illinois; and [3]VIMSS (Virtual Institute of Microbial Stress and Survival) http://vimss.lbl.gov/

Project Goals: See goals for abstract 36.

Because one of the central aims of the 'Environmental Stress Pathway Project' is to elucidate regulatory networks critical to processes of interest to the DOE, the Computational Core has garnered valuable transcriptional and proteomic profiles under various environmental stress conditions. This data has been essential to enhancing our biological systems knowledge of the model metal reducer *Desulfovibrio vulgaris*. To further understand how this organism and its relatives compete in the environment and regulate their metabolism in contaminated sites, additional studies on intricate regulatory cascades are imperative. One potential alternative regulatory mechanism currently being targeted by the ESPP is that of small non-coding RNA molecules (sRNAs). Ranging in size from 20-200 nucleotides (nt), sRNAs predominantly affect gene regulation by binding to complementary mRNA in an anti-sense fashion and therefore provide an immediate regulatory response independent of protein modification. Whereas data are available for sRNAs in such prokaryotes as *Escherichia coli*, *Archaeoglobus fulgidus*, *Pseudomonas aeruginosa*, and *Vibrio* species, no information is available on metal-reducing bacteria, or for that matter, members of the delta-*proteobacteria*.

* Presenting author

In an effort to identify sRNAs in *D. vulgaris*, a strategy for cloning total RNA ranging in size from 20-200 nt was employed. Following addition of directional aptamer sequences, cDNAs were produced and cloned for sequencing. Sequence analysis of a small portion of the resulting cDNA library yielded two identical ~65 nt sRNA clones (Dv-sRNA2) possessing complementary sequence to the RBS of open reading frame (ORF) DVU0678. While DVU0678 is adjacent to the Dv-sRNA2 gene, the ORF is transcribed from the opposite chromosomal strand. Northern analysis specialized for sRNAs verified the expression of Dv-sRNA2 as an individual transcript under anaerobic lactate/sulfate growth (LS4D medium). These data suggest that when Dv-sRNA2 is transcribed, translation of DVU0678 will be inhibited. DVU0678 has been annotated to encode a putative 34 amino acid protein unique to *D. vulgaris* strains Hildenborough and DP4, hampering our abilities to discern the role of DVU0678 in the cell. Further sequence analysis of the Dv-sRNA DNA locus by 'PromScan' identified a putative sigma^{54}-recognition site (97% probability) 43 nt upstream of the predicted sRNA transcriptional start site and therefore suggests that Dv-sRNA2 may be member of the sigma^{54} regulon. A perfect stem-loop terminator was also identified 26 nt downstream of the Dv-sRNA2 DNA sequence. Current analysis is underway to ascertain the expression profile for this sRNA as well as the effect over-expression has on the physiology and transcriptional response of *D. vulgaris* under multiple environmental conditions.

50

GTL

The Dynamics and Genetic Adaptation to Salt Stress in Experimental Evolution of *Desulfovibrio Vulgaris* Hildenborough

Aifen Zhou,[1,6]* Zhili He,[1,6] Marcin P. Joachimiak,[2,6] Paramvir S. Dehal,[2,6] **Adam P. Arkin**,[2,6] Kristina Hillesland,[4,6] **David Stahl**,[4,6] **Judy Wall**,[5,6] **Terry C. Hazen**,[3,6] and **Jizhong Zhou**[1,6] (jzhou@ou.edu)

[1]Institute for Environmental Genomics, Department of Botany and Microbiology, University of Oklahoma, Norman, Oklahoma; [2]Physical Biosciences Division and [3]Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, California; [4]Civil and Environmental Engineering, University of Washington, Seattle, Washington; [5]Departments of Biochemistry and Molecular Microbiology and Immunology, University of Missouri, Columbia, Missouri; and [6]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

Project Goals: See goals for abstract 36.

One of the greatest challenges in biology is to understand how genotype and environment interact to determine the phenotype and fitness of an organism. With the recent advances in genome sequencing and high-throughput genomic technologies, it becomes capable to link sub-cellular molecular/metabolic processes with the population-level processes, functions and evolution. One of our goals of the new proposal is to bring the environmental microbe, *D. vulgaris* Hildenborough to the model organism status (Aim 1). This study particularly intends to mimic the environmental conditions (salt stress) to address the evolution of *DvH* under such conditions in the lab. Such a study is expected to generate different DvH strains, and allows us to identify multiple beneficial mutations for salt adaptation. Therefore, this study directly links stress responses to the evolution of DvH, and will provide information for our integrative understanding of gene function, regulation, networking and evolution of DvH. To determine the long-term evolutionary responses, diversifications and adaptation of DvH to environmental stresses, the control and evolved cell lines (6 lines each) were obtained from a single DvH colony directly from the original glycerol stock. LS4D was used as standard culture medium for the control lines. Evolved lines were cultured on LS4D + 100 mM NaCl. Cells were kept at 37°C and transferred every 48 hrs with one to one hundred dilutions. The cells from every 100 generations were archived. The results demonstrated that the adaptation of DvH to salt stress was a dynamical process. The enhanced salt tolerance to higher salt (LS4D + 250 mM NaCl) of evolved lines was observed at 300 generations; and this phenomenon became more and more obvious with the increase of generations. Compared to the ancestor and paralleled control lines, both the growth rate and final biomass of the evolved lines were higher. The de-adaptation experiment on 1000 generation evolved lines provided the evidence that the phenotype was due to the genetic change instead of physiological adaptation. The gene expression profile of the 1000 generation evolved lines showed that some poly-cistronic operons such as hmcF-E-D-C-B-A (functional genes), rrf2-rrf1 (regulatory genes), LysA-2-LysX (functional genes) and DVU3290-3291-3292 (glutamate synthase) were significantly up-regulated compared to the control lines. Next, de-adaptation experiment need to be done on different generation samples to confirm that the beneficial genetic mutations are stable and genome sequencing will be performed to reveal the possible genetic mutations.

* Presenting author

51

GTL

Applications of GeoChip to Examine Functional Microbial Communities in Metal Contaminated Environments

Joy D. Van Nostrand,[1,8]* Liyou Wu,[1] Sanghoon Kang,[1] Patricia Waldron,[1] Yuting Liang,[1] Ye Deng,[1] Zhili He,[1,8] Weimin Wu,[2] Heidi Gough,[3] Sue Carroll,[4] Chris Schadt,[4,8] Anthony Palumbo,[4,8] Dave Watson,[4] Craig Criddle,[2] Phil Jardine,[4] Brett Baldwin,[5] Aaron Peacock,[5] Phil Long,[6] David Stahl,[3,8] **Terry C. Hazen**,[7,8] and **Jizhong Z. Zhou**[1,8] (jzhou@ou.edu)

[1]University of Oklahoma, Norman, Oklahoma; [2]Stanford University, Stanford, California; [3]University of Washington, Seattle, Washington; [4]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [5]University of Tennessee, Knoxville, Tennessee; [6]Pacific Northwest National Laboratory, Richland, Washington; [7]Lawrence Berkeley National Laboratory, Berkeley, California; and [8]VIMSS (Virtual Institute of Microbial Stress and Survival) http://vimss.lbl.gov/

Project Goals: See goals for abstract 36.

The GeoChip 2.0, a functional gene microarray, allows for the simultaneous detection of >10,000 genes involved in the geochemical cycling of C, N, and S, metal reduction and resistance, and organic contaminant degradation. The GeoChip has been used to examine dynamic functional and structural changes in microbial communities from many different environments. Here, five examples of studies utilizing the GeoChip to examine microbial communities at metal contaminated sites are presented. These studies specifically address and support Aim 3's goal to 'examine the natural community dynamics in the field for correlation to laboratory observations'. These initial studies illustrate the ability of the GeoChip to provide direct linkages between microbial genes/populations and ecosystem processes and functions. The first three studies examined areas within the U.S. DOE's Field Research Center (FRC) in Oak Ridge, TN. (1) Microbial communities within a pilot-scale test system established for the biostimulation of U(VI) reduction in the subsurface by injection of ethanol were examined. The microarray data indicated that during the U(VI) reduction period, both FeRB and SRB populations reached their highest levels at Day 212, followed by a gradual decrease over the following 500 days. The U concentrations in the groundwater were significantly correlated with the total abundance of c-type cytochrome genes and with the total abundance of *dsrAB* (dissimilatory sulfite reductase) genes. Mantel test analysis of microarray and chemical data indicated a significant correlation between the U concentration and total c-cytochrome or *dsrAB* gene abundance. Changes in more than a dozen individual c-type cytochrome genes and more than 10 *dsrAB*-containing populations showed significant correlations to the changes in U concentration among different time points, indicating their importance in uranium reduction. (2) In a different study of the same system, the effects of dissolved oxygen (DO) and ethanol on the stability of the bioreduced area were examined. Canonical correspondence analysis (CCA) and Mantel test analysis revealed that ethanol and sulfide concentrations showed the greatest correlation to the functional community structure. Detrended correspondence analysis (DCA) showed a shift towards a different community structure after ethanol injections resumed compared to the periods of starvation and exposure to DO. Changes in the functional community structure were similar in both wells; however, the community in FW101-2 was more affected by DO than in FW102-3. This is most likely because FW101-2, located closest to the injection wells, had a greater increase in DO than FW102-3, located further from the injection well. Hierarchical clustering showed that cytochrome c genes grouped based on DO exposure, starvation, or ethanol addition, while dissimilatory sulfite reductase (*dsr*) genes grouped only by starvation or ethanol addition. However, when DO levels increased, the relative abundance of *dsr* genes decreased while cytochrome genes seemed unaffected. Overall, results indicated that ethanol was the main factor affecting community structure, although some changes could be attributed to DO. (3) In the third study from the FRC, analysis of groundwater monitoring wells along a contamination gradient revealed less overlap between wells with different levels of U and NO_3 contamination. While diversity of nitrate-fixation genes decreased in NO_3 -contaminated wells, the diversity of metal reduction and resistance genes did not correlate with metal concentrations. Signal intensity did, however, increase in heavily contaminated wells, indicating a larger percentage of organisms with metal- related genes. Sulfate-reduction genes had greater diversity and greater signal intensity in more contaminated wells. Individual principle component analyses (PCA) of the gene diversity and geochemistry of the wells separated them in similar ways. CCA indicated that pH was an important variable that correlated with gene diversity in the lowest-contamination well, while NO_3 and U correlated with the most highly contaminated well. Overall, contaminant level appears to have significant effects on the functional gene diversity along the contaminant plume at the FRC. (4) We have also used GeoChip to examine a Uranium Mill Tailings Remedial Action (UMTRA) site (Rifle, CO). Two adjacent mini-galleries were driven to Fe-reducing and

* Presenting author

SO_4-reducing conditions, respectively, in order to better understand the long-lived U (VI) loss and to constrain the relative impacts of SO_4 and Fe (III) reduction. Cluster analysis results showed samples in the same locations grouped together, regardless of geochemistry. The *dsr* genes increased when conditions were driven to sulfate-reduction. DCAs of both the functional community structure and environmental conditions (Fe^{2+}, H_2S, DO, pH, conductivity, potential and Eh) showed background, Fe-reducing, and SO_4-reducing samples clustered together, respectively. CCA of environmental parameters and functional genes indicated Fe^{2+} was the most significant geochemical variable for community structure. (5) Additionally, metal contaminated freshwater lake sediments (Lake DePue, IL) were analyzed to examine the link between functional genes and the environmental gradient of metal contamination. Based on non-metric multidimensional scaling (NMDS), the microbial communities were separated based on sampling regions. In addition, there were different groupings between samples with the highest levels of contamination and the lesser contaminated samples. These studies demonstrate the analytical power of the GeoChip in examining microbial communities. This is the first comprehensive microarray available for studying the functional and biogeochemical cycling potential of microbial communities.

52

GTL

Further Development and Applications of GeoChip 3.0 for Microbial Community Analysis

Zhili He,[1,5]* Ye Deng,[1,5] Joy D. Van Nostrand,[1,5] Liyou Wu,[1,5] Christopher Hemme,[1,5] Terry J. Gentry,[2] **Adam P. Arkin**,[3,5] **Terry C. Hazen**,[4,5] and **Jizhong Zhou**[1,5] (jzhou@ou.edu)

[1]University of Oklahoma, Norman, Oklahoma; [2]Texas A&M University, College Station, Texas; [3]Physical BioScience Division, Lawrence Berkeley National Laboratory, Berkeley, California; [4]Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, California; and [5]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

Project Goals: See goals for abstract 36.

Microarrays fabricated with the genes encoding key, functional enzymes involved in various biological and geochemical cycling processes are referred to as functional gene arrays (FGAs). On the basis of GeoChip 2.0, which contains 24,243 oligonucleotide (50mer) probes and covers > 10,000 gene sequences in >150 functional groups involved in nitrogen, carbon, sulfur and phosphorus cycling, metal reduction and resistance, and organic contaminant degradation, a new generation of GeoChip (GeoChip 3.0) has been developed. GeoChip 3.0 has several new features compared to GeoChip 2.0. First, GeoChip 3.0 is expected to cover >37,000 gene sequences for more than 300 gene families, and such a coverage allows us to obtain more information about microbial communities and analyze more diverse environmental samples. Second, the homology of automatically retrieved sequences by key words is verified by HUMMER using seed sequences so that unrelated sequences are removed. Third, GeoChip 3.0 includes phylogenic markers, such as *gyrB.* Fourth, a software package (including databases) has been developed for sequence retrieval, probe and array design, probe verification, array construction, array data analysis, information storage, and automatic update, which greatly facilitate the management of such complicated array, especially for future update. Fifth, a universal standard has been implemented in GeoChip 3.0 so that data normalization and comparison can be conducted. Sixth, a genomic standard is also used to quantitatively analyze gene abundance. Finally, GeoChip 3.0 also includes GeoChip 2.0 probes, and those GeoChip 2.0 probes are checked against new databases. Disqualified probes are flagged. GeoChip 3.0 will provide more capability for studying biogeochemical processes and functional activities of microbial communities important to human health, agriculture, energy, global climate change, ecosystem management, and environmental cleanup and restoration. It is also particularly useful for providing direct linkages of microbial genes/populations to ecosystem processes and functions. In the new proposal, we will explore to discover the molecular mechanisms by which microbial community structure, function, and stability affects stress response and activity in the lab and in the field. GeoChip 3.0 or similar but more specific functional gene arrays will directly examine the natural community dynamics in the field for correlation to laboratory observations (Aim 3 of the new proposal), and will facilitate our understanding of the molecular basis of microbial community formation, stability, resilience, specificity, activity and ecology. In addition, all GeoChip data can be integrated to trace the path from genome to community phenotypes.

* Presenting author

53

GTL

Comparative Metagenomics of Microbial Communities from Pristine and Contaminated Groundwater

Christopher L. Hemme,[1,2]* Ye Deng,[1] Terry J. Gentry,[3] Liyou Wu,[1,2] **Matthew W. Fields**,[2,4] Susannah Green-Tringe,[5] Chris Detter,[5] Kerrie Barry,[5] Nikos Kyrpides,[5] David Watson,[6] Paul Richardson,[5] **Terry C. Hazen**,[2,7] James Tiedje,[8] Eddy Rubin,[5] and **Jizhong Zhou**[1,2] (jzhou@ou.edu)

[1]Institute for Environmental Genomics, University of Oklahoma, Norman, Oklahoma; [2]Virtual Institute of Microbial Stress and Survival http://vimss.lbl.gov/; [3]Department Soil and Crop Sciences, Texas A&M University, College Station, Texas; [4]Department of Microbiology, Montana State University, Bozeman, Montana; [5]DOE Joint Genome Institute, Walnut Creek, California; [6]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [7]Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, California; and [8]Center for Microbial Ecology, Michigan State University, East Lansing, Michigan

Project Goals: See goals for abstract 36.

To better understand the evolutionary and physiological mechanisms by which microbial communities adapt to environmental stress, community-scale genomic analyses are necessary. To this end, microbial community DNA isolated from contaminated groundwater located at the U.S. Dept. of Energy Field Research Center (FRC) (Area 3, Well FW106) in Oak Ridge, TN, was analyzed to determine the effects of chronic exposure to multiple stressors (including low pH, nitrate, uranium and organic solvents) on the nascent microbial community structure. Consistent with previous 16S rRNA gene surveys conducted at the FRC, metagenomic analysis indicates a very low diversity community (~13 OTU) dominated by clonal denitrifying γ- and β-proteobacterial populations. Metabolic reconstruction of the dominant γ-proteobacterial species reveals adaptations for specific geochemical parameters including the following: denitrification and ammonium assimilation pathways; mechanisms for resistance to low pH; pathways for degradation of organic compounds such as1,2-dichloroethene, acetone, butanol, methanol and formaldehyde; accumulation of multiple heavy metal efflux systems (*czcABC*, *czcD*, *cadA*-family, *mer* operon genes, etc.). Analysis indicates that lateral gene transfer is the predominant mechanism of introducing genetic variation into the community, resulting in the lateral acquisition of geochemical resistance genes (e.g. acetone carboxylation, heavy metal efflux systems etc.). A particularly interesting adaptation is the number and variety of toxic efflux genes identified in the FW106 metagenome. In particular, genes encoding NarK nitrate/nitrite antiporters, CzcABC divalent cation transports and CzcD divalent cation transporters are highly abundant, suggesting that these genes are important in community stress response. The most likely physiological effect of this accumulation of transporters would be to increase the baseline rate of toxic efflux from the cell. Rapid efflux of toxins in conjunction with specific degradation and detoxification pathways, thus appears to be a major survival strategy under stressed conditions.

The FW106 sample was compared to a second groundwater metagenome from a pristine FRC site (FW301) to determine differences between the two communities. In contrast to the low species diversity of FW106, the FW301 is represented by multiple phyla including all 5 classes of proteobacteria, *Planctomycetes*, *Chloriflexi*, *Actinobacteria*, *Acidobacteria*, *Bacteroidetes* and *Firmicutes*. In contrast to the FW106 sample which resulted in significant read assembly, the FW301 sample is composed largely of single reads that do not assemble into contigs (95%). Interestingly, most of the geochemical resistance mechanisms identified in FW106 are also present in FW301 as well as other resistance mechanisms (e.g. aromatic degradation) that are not found in FW106. Abundance profiling of geochemical and cytochrome genes between FW106 and FW301 and between FW106 and the acid mine drainage (AMD) metagenome show distinct environmental signatures between the samples, including the accumulation of cytochrome *c*553 genes in the FW106 metagenome which may play a role in metal resistance. The AMD metagenome also showed an abundance of CzcABC genes, suggesting that accumulation of these genes may be an important resistance mechanism in metal-stressed communities. The identification of FW106 stress response genes will also permit more in-depth studies of the dynamics of stressed FRC communities using functional genomics tools such as the GeoChip.

* Presenting author

54

GTL

Comparative Genomics of Ethanol-Producing Thermoanaerobacter Species

Christopher L. Hemme,[1,7]* **Matthew W. Fields**,[2,7] Qiang He,[3] Zhiguo Fang,[1] Zhili He,[1,7] J.C. Detter,[4] Kerrie Barry,[4] Miriam Land,[4] Loren Hauser,[4] Alla Lapidus,[5] Cliff S. Han,[6] Paul Richardson,[5] Eddy Rubin,[6] and **Jizhong Zhou**[1,7] (jzhou@ou.edu)

[1]Institute for Environmental Genomics, Department of Botany and Microbiology, University of Oklahoma, Norman, Oklahoma; [2]Department of Microbiology, Montana State University, Bozeman, Montana; [3]Department of Civil and Environmental Engineering, Temple University, Philadelphia, Pennsylvania; [4]Genome Analysis and Systems Modeling Group, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [5]DOE Joint Genome Institute, Walnut Creek, California; [6]Los Alamos National Laboratory, Los Alamos, New Mexico; and [7]VIMSS (Virtual Institute of Microbial Stress and Survival) http://vimss.lbl.gov/

Project Goals: See goals for abstract 36.

Recent global fluctuations in the supply and demand of petroleum have prompted a call for increased research into biologically-derived fuels (biofuels). Of particular interest are those processes involving production of biofuels from cellulosic biomass. The conversion of biomass to biofuels is critically dependent on the flux of carbon through the fermentation pathways. To better understand the carbon pathways involved in this process, the genomes of two strains of ethanol-producing *Thermoanaerobacter* species (*T. pseudoethanolicus* 39E and *T. ethanolicus* X514) have been sequenced and compared to a third species, *Caldanaerobacter subterraneus* supsp. *tengcongensis* (formaerly *Thermoanaerobacter tengcongensis*). Strain 39E is a well-characterized strain isolated from a Yellowstone hot spring and is capable of degrading xylan and starch to ethanol at high yields. Strain X514 is a metal-reducing bacteria isolated from the deep subsurface and is predicted to have been geographically isolated from 39E for ~250 MY. Both strains show a high degree of gene conservation but marked genomic rearrangements compared to each other and to *C. subterraneus*. The common *Thermoananerobacter* ancestor underwent a genome rearrangement around the origin of replication following divergence from *C. subterraneus*. Additional large-scale rearrangements occurred in X514 following its divergence from 39E which has resulted in a strong strand asymmetry in X514, with 62.3% of the X514 genes being located on the leading strand. A survey of unique genes between the strains reveals lineage-specific gene expansions in the two strains including individual unique sugar transporter profiles and an increased number of P-type metal translocating ATPase genes in X514. Several of these gene clusters are associated with phage genes, suggesting phage-mediated lateral gene transfer. The activity of phage in the X514 environment is also evidenced by a significantly expanded CRISPR region in X514 compared to 39E, which implies phage resistance activities. X514 also encodes a complete Vitamin B_{12} biosynthesis pathway that is partially lacking in 39E and *C. subterraneus*. This pathway is associated with a large genomic rearrangement observed between the two *Thermoanaerobacter* strains. Exogenous B_{12} has been shown to increase ethanol yields in certain strains of *Clostridium thermocellum* and thus the ability to synthesize B_{12} *de novo* may contribute to the increased ethanol yields observed with X514.

Comparisons of ethanol yields of 39E and X514 in both mono- and batch culture suggest that X514 is a more efficient ethanol producer than 39E. Because X514 was isolated from a presumably nutrient poor environment, it is hypothesized that the strain has evolved high-affinity carbon scavenging mechanisms that may contribute to the observed high ethanol yields. Such mechanisms may manifest as novel genes or in modifications to existing genes. Metabolic reconstruction of the *Thermoanaerobacter* species reveals insights into carbon metabolism and niche adaptation of the two strains. Both strains are capable of metabolizing glucose and xylan to ethanol with a novel bifunctional secondary alcohol dehydrogenase serving as the terminal enzyme in the pathway. Differences are noted in the carbon metabolism pathways of the two strains, including a complete KDPG metabolism pathway in 39E and the lack of a complete methylglyoxal shunt in X514. Abundance profiles of leading vs lagging strand genes based on COG categories for both 39E and X514 show an overabundance of carbon transport and metabolism genes on the leading strand of X514 compared to the lagging strand of X514 and the leading strand of 39E, suggesting that carbon metabolism genes may be more highly expressed in X514. However, an initial analysis of codon adaptation indices (CAI), a possible measure of gene expression, does not support this hypothesis. CAI analysis, however, does show that cellobiose-specific PTS component genes are highly expressed in both strains which may contribute to the increased ethanol production of these strains in coculture with cellulosic bacteria.

* Presenting author

To complement this research, a request for sequencing the genomes of an additional 20 ethanol-producing *Clostridia* strains has been approved by JGI. Strains were chosen from among the genera *Clostridium*, *Thermoanaerobacter*, *Thermoanaerobacterium* and *Acetivibrio* based on prior knowledge, phylogeny, unique physiology and industrial applications. The expansion of the genomic database of industrially-important *Clostridia* is expected to provide substantial benefits in the understanding of this class of organisms. Additional computational and functional genomic analyses are currently being conducted to further elucidate the mechanisms of carbon usage and ethanol production in these bacteria and to correlate the computational predictions with experimental gene expression data. A better understanding of the expression of biomass conversion genes will aid in future efforts to engineer these bacteria for industrial use.

55

GTL

Metabolic Flux Analysis of *Shewanella spp* Central Carbon Metabolism Reveals Evolutionary, Genetic, and Environmental Robustness

Yinjie J. Tang,[1,2]* Hector Garcia Martin,[3] Paramvir S. Dehal,[1,2] Adam Deutschbauer,[1,2] Xavier Llora,[4] Adam Meadows,[5] **Adam P. Arkin**[1,2,6] (aparkin@lbl.gov), and **Jay D. Keasling**[1,2,6,7]

[1]Virtual Institute of Microbial Stress and Survival http://vimss.lbl.gov, Lawrence Berkeley National Laboratory, Berkeley, California; [2]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; [3]Microbial Ecology Program, DOE Joint Genome Institute, Walnut Creek, California; [4]National Center for Supercomputing Applications, University of Illinois, Urbana-Champaign, Illinois; [5]Genentech, Inc., South San Francisco, California; [6]Department of Bioengineering, University of California, Berkeley, California; and [7]Department of Chemical Engineering, University of California, Berkeley, California

Project Goals: See goals for abstract 36.

A major goal of the Environmental Stress Pathway Project (ESPP) is to uncover the mechanisms that allow microbes to sustain activity under the environmental stress seen in field settings. In this study, we used ^{13}C isotopomer analysis to elucidate the metabolic flux through central metabolism of eight *Shewanella* species. Metabolite fluxes link genes, proteins and metabolites to macroscopic biological functions. In spite of its importance, only a few, not thoroughly tested, general principles have been proposed to predict and understand the flux configuration of an organism. Among those general principles, robustness of central metabolism has been reported with respect to genetic perturbation. Here we show that the relative metabolic flux distributions are very similar for phylogenetically and environmentally diverse members of the *Shewanella* genus. This phylogenetic robustness suggests understanding microbial fluxomics in terms of metabolic types (or metabotypes), as opposed to phylotypes. In addition to phylogenetic, environmental and genetic robustness our data shows flexibility in the relative flux profiles when adapting to different carbon sources.

56

GTL

VIMSS Computational Core

Paramvir S. Dehal[1,2]* (PSDehal@lbl.gov), Eric J. Alm,[1,3] Dylan Chivian,[1,2] Katherine H. Huang,[1,2] Y. Wayne Huang,[1,2] Marcin P. Joachimiak,[1,2] Keith Keller,[1,4] Morgan N. Price,[1,2] and **Adam P. Arkin**[1,2,4,5,6]

[1]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov/; [2]Lawrence Berkeley National Laboratory, Berkeley, California; [3]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts; [4]University of California, Berkeley, California; [5]Howard Hughes Medical Institute, Chevy Chase, Maryland; and [6]Department of Bioengineering, University of California, Berkeley, California

Project Goals: See goals for abstract 36.

Background: The VIMSS Computational Core group is responsible for data management, data integration, data analysis, and comparative and evolutionary genomic analysis of the data for the VIMSS project. We have expanded and extended our existing tools sets for comparative and evolutionary genomics and microarray analysis as well as creating new tools for our proteomic and metabolomic data sets. Our analysis has been incorporated into our comparative genomics website MicrobesOnline (http://www.microbesonline.org) and made available to the wider research community. By taking advantage of the diverse functional and comparative datasets, we have been able to pursue large evolutionary studies.

Data Analysis: During the course of analysis of various stress responses of DvH, the computational core has

* Presenting author

continued to develop new statistical analyses of data that take advantage of the predicted regulatory structures (operons, regulons, etc.) from our comparative analyses. This year we have used these analyses to investigate the response of DvH to oxygen stress and pH stress. Our analysis has focused on the combined results from both transcriptomic and proteomic datasets to interpret oxygen stress. Additionally, we have worked with metabolomic datasets within the framework of predicted metabolic activities to find missing pathway members.

Data Management: All data generated by ESPP continues to be stored in our Experimental Information and Data Repository (http://vimss.lbl.gov/EIDR/). Researches have access to datasets from biomass production, growth curves, image data, mass spec data, phenotype microarray data and transcriptomic, proteomic and metabolomic data. New functionality has been added for storage of information relating to mutants and protein complex data, in addition to new visualization for assessing existing data sets such as the phenotype microarrays.

The MicrobesOnline Database: The MicrobesOnline database (http://www.microbesonline.org) currently holds over 700 microbial genomes and will be updated quarterly, providing an important comparative genomics resource to the community. New functionality added this year includes the addition of a thousands of phage genomes and plasmids, an updated user interface for the phylogenetic tree based genome browser that allows users to view their genes and genomes of interest within an evolutionary framework, tools to compare multiple microarray expression data across genes and genomes, addition of external microarray data from the Many Microbial Microarrays Database, integration with the RegTransBase of experimentally verified regulatory binding sites and links to three dimensional protein structures of proteins and their close relatives.

MicrobesOnline continues to provide an interface for genome annotation, which like all the tools reported here, is freely available to the scientific community. To keep up with the rapidly expanding set of sequenced genomes, we have begun to investigate methods for accelerating our annotation pipeline. In particular we have completed work on methods to speed up the most time consuming process, homology searching through HMM alignments and all against all BLAST. These methods now enable us to deal with the many millions of gene sequences generated from metagenomics.

Over the next year, several new features will be added to the MicrobesOnline resource. Microarray expression data will be added from the NCBI GEO database, in addition to datasets generated from the VIMSS team. To supplement the analysis tools we already have, enrichment of functional genes and operon-wise analysis, we will provide tools for comparing multiple experiments across multiple genomes. We will also expand our regulatory binding motif search to incorporate co-expression data to support predictions.

Evolutionary Analysis: The computational core continues work on understanding the evolution of regulatory networks. Transcription factors form large paralogous families and have complex evolutionary histories. Our analysis shows that putative orthology derived from bidirectional best hits across distantly related bacteria are usually not true evolutionary orthologs. Additionally, these false orthologs usually respond to different signals and regulate distinct pathways. Even in more closely related genomes, such as *E. coli* and *Shewanella oneidensis*, bidirectional best hits have a high error rate. By studying transcription factors with phylogenetic trees, we show that through the use of gene-regulon correlations, together with sequence analysis of promoter regions for confirmation, bacterial regulatory networks may evolve more rapidly than previously thought.

57

GTL

MicrobesOnline: An Integrated Portal for Comparative Microbial Functional Genomics

Marcin P. Joachimiak,[1,2]* Y. Wayne Huang,[1,2] Katherine H. Huang,[1,2] **Eric J. Alm**,[1,3] Dylan Chivian,[1,2] Paramvir S. Dehal,[1,2] Keith Keller,[1,4] Morgan N. Price,[1,2] and **Adam P. Arkin**[1,2,4,5] (aparkin@lbl.gov)

[1]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov/; [2]Lawrence Berkeley National Laboratory, Berkeley, California; [3]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts; [4]University of California, Berkeley, California; and [5]Department of Bioengineering, University of California, Berkeley, California

Project Goals: See goals for abstract 36.

The Virtual Institute for Microbial Stress and Survival (VIMSS, http://vimss.lbl.gov/) funded by the Dept. of Energy's Genomics:GTL Program, is dedicated to using integrated environmental, functional genomic, and comparative sequence and phylogeny data to understand mechanisms by which microbes and microbial communi-

* Presenting author

ties survive in uncertain environments while carrying out processes of interest for bioremediation and energy generation. To support this work, VIMSS has developed a Web portal, along with computational analysis and an underlying database, for comparative functional genomics of bacteria and archaea. MicrobesOnline (http://www.microbesonline.org) has been enabling comparative genome analysis since 2003. The portal currently includes 702 complete microbial genomes (662 bacterial, 40 archaeal, 1770 viruses, 268 plasmids) and offers a suite of analysis and tools including: comprehensive gene family and domain annotations, information on three dimensional structure representatives and similarities, an interactive gene expression heatmap browser along with functionality for gene expression profile searches, a multi-species genome browser, operon and regulon predictions, a combined gene and species phylogeny browser, a gene ontology browser, a metabolic pathway browser, a workbench for sequence analysis (including sequence motif detection, motif searches, sequence alignment and phylogeny reconstruction), and capabilities for community annotation of genomes.

VIMSS integrates functional genomic data and provides novel web-based viewing and analysis tools for gene expression microarray, proteomic, metabolomic, and phenotype microarray data. Currently, these data are mostly project generated for wild-type and mutants *of Desulfovibrio vulgaris* and *Shewanella oneidensis* exposed to stress conditions found at DOE field sites. However, the organism scope is being expanded, and recently the *E. coli* gene expression data compendium from the Many Microbes Microarray Database (M3D) has been imported into MicrobesOnline. Additionally VIMSS has developed capabilities to analyze microarray experiments performed on multiple species simultaneously. Selecting an organism or gene of interest in MicrobesOnline leads to information about and data viewers for VIMSS experiments conducted on that organism and involving that gene or gene product. It is possible to view microarray data from multiple stress conditions as an interactive heatmap and to analyze correlations between gene expression results from different experiments. Among the major new features is the ability to search the microarray data compendium for genes with gene expression profiles similar to a query expression profile (either based on a gene or set of genes). Such new compendium-wide functionalities allow to observe patterns in gene expression changes across multiple conditions and genes, and to search for similarities to these patterns. The information integration and analysis performed by VIMSS serves not only to generate insights into the stress responses and their regulation in these microorganisms, but also to document VIMSS experiments, allow contextual access to experimental data, and facilitate the planning of future experiments. VIMSS also is incorporating into MicrobesOnline publicly available functional genomics data from published research, so as to centralize and synergize data on and analysis of microbial physiology and ecology in a unified comparative functional genomic framework.

58

GTL

The Analysis and Expansion of Regulatory Binding Site Data in a Wide Range of Bacteria Using a Semi-Automatic System—RegTransBase

Michael J. Cipriano,[1] Alexei E. Kazakov,[2] Dmitry Ravcheev,[2] **Adam P. Arkin**[1,3,4,5] (aparkin@lbl.gov), Mikhail S. Gelfand,[2,6] and **Inna Dubchak**[1,5,7*]

[1]Lawrence Berkeley National Laboratory, Berkeley, California; [2]Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia; [3]Howard Hughes Medical Institute, Chevy Chase, Maryland; [4]Department of Bioengineering, University of California, Berkeley, California; [5]Virtual Institute of Microbial Stress and Survival http://vimss.lbl.gov/; [6]Faculty of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia; and [7]DOE Joint Genome Institute, Walnut Creek, California

Project Goals: See goals for abstract 36.

RegTransBase, a database describing regulatory interactions in prokaryotes, has been developed as a component of the MicrobesOnline/RegTransBase framework successfully used for interpretation of microbial stress response and metal reduction pathways. It is manually curated and based on published scientific literature. RegTransBase describes a large number of regulatory interactions and contains experimental data which investigates regulation with known elements. It is available at http://regtransbase.lbl.gov. Currently, the database content is derived from more than 4000 relevant articles describing over 9000 experiments in relation to 155 microbes. It contains data on the regulation of ~14000 genes and evidence for ~7500 interactions with ~850 regulators.

RegTransBase additionally provides an expertly curated library of alignments of known transcription factor binding sites covering a wide range of bacterial species. Each alignment contains information as to the transcription factor which binds the DNA sequence, the exact location

* Presenting author

of the binding site on a published genome, and links to published articles.

RegTransBase builds upon these alignments by containing a set of computational modules for the comparative analysis of regulons among related organisms. These modules guide a user through the appropriate steps of transferring known or high confidence regulatory binding site results to other microbial organisms, allowing them to study many organisms at one time, while warning of analysis possibly producing low confidence results, and providing them with sound default parameters.

There is an increasingly tight coupling of RegTransBase with MicrobesOnline in reporting cis-regulatory sites and regulatory interactions, and integrating RegTransBase searches into MicrobesOnLine cart functions.

59

GTL

Evolutionary History of Gene Regulation in Bacteria

Morgan N. Price,[1,2] Paramvir S. Dehal,[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, California; and [3]Department of Bioengineering, University of California, Berkeley, California

Project Goals: See goals for abstract 36.

Analysis of gene regulation in ESPP bacteria relies on comparisons to model organisms, and hence on assumptions about how gene regulation evolves. To test these assumptions we examined the evolutionary histories of transcription factors and of regulatory interactions from the model bacterium *Escherichia coli* K12. We show that although most transcription factors have paralogs, these usually arose by horizontal gene transfer rather than by duplication within the *E. coli* lineage, as previously believed. Most neighbor regulators — regulators that are adjacent to genes that they regulate — were acquired by horizontal gene transfer, while most global regulators evolved vertically within the gamma-Proteobacteria. Neighbor regulators are often acquired together with the operon that they regulate, which suggests that the proximity is maintained by repeated transfers, and also aids the prediction of the regulators' function. Because of the complex evolutionary histories of most transcription factors, bidirectional best hits tend to be misleading, and most annotations of bacterial regulators are probably incorrect.

When we analyzed the histories of regulatory interactions, we found that the evolution of regulation by duplication was rare, and surprisingly, many of the regulatory interactions that are shared between paralogs result from convergent evolution. Furthermore, horizontally transferred genes are more likely than other genes to be regulated by multiple regulators, and most of this complex regulation probably evolved after the transfer. Finally, gene regulation is often not conserved, even within the gamma-Proteobacteria. Our results suggest that the bacterial regulatory network is evolving rapidly under positive selection. Such rapid rewiring of gene regulation may be crucial for adaptation to new niches.

60

GTL

MicroCOSM: Phylogenetic Classification of Metagenomic Data Using Microbial Clade-Oriented Sequence Markers

Dylan Chivian[1,2]* (DCChivian@lbl.gov), Paramvir S. Dehal,[1,2] and **Adam P. Arkin**[1,2,3]

[1]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov; [2]Lawrence Berkeley National Laboratory, Berkeley, California; and [3]University of California, Berkeley, California

Project Goals: See goals for abstract 36.

The VIMSS/ESPP2 project requires understanding of the microbial communities at contaminated field sites and, among other methods, will employ metagenomics in this endeavor. Metagenomics projects that seek to elucidate the population structure of microbial ecosystems are faced with the related computational challenges of classifying the sequences obtained and quantifying which organisms are present within a sample. Individually low-proportion species usually make up a large fraction of microbial communities, complicating their classification and quantification using traditional phylogenetic marker approaches. Such species usually don't yield sufficient read depth to assemble into longer sequences, leaving fragments that rarely contain traditional markers such as the small subunit (SSU) rRNA gene. BLAST-based approaches for analysis of metagenomic sequences [1] compensate for this rarity of traditional markers, but may be confounded by genes that are subject to horizontal transfer or duplication. Another approach instead makes

* Presenting author

use only of reliable non-transferred single-copy genes [2] to classify and quantify the organisms present within a sample, but the application has so far been limited to the use of a fairly small set of universal genes found in all organisms. In this work, we have extended the latter approach, boosting the set of reliable marker genes from only about 30-40 universal genes to several hundred by identifying sets of single-copy genes that are not subject to inter-clade horizontal transfer through investigation of finished bacterial and archaeal genomes. These clade-oriented sequence markers allow for a method, which we have named "MicroCOSM", that greatly increases the probability that a marker will be found in any given sequence and therefore offers improved coverage for phylogenetic classification and quantification of microbial types in an environmental sample.

References

1. Huson D.H., Auch A.F., Qi J., Schuster S.C. (2007) "MEGAN analysis of metagenomic data." *Genome Res.* **17**(3):377-86.
2. von Mering C., Hugenholtz P., Raes J., Tringe S.G., Doerks T., Jensen L.J., Ward N., Bork P. (2007) "Quantitative phylogenetic assessment of microbial communities in diverse environments." *Science* **315**(5815):1126-30.

61

GTL

AdaptML: A Maximum Likelihood Framework for Integrating Evolutionary and Ecological Reconstructions

Lawrence David,[1] Dana Hunt,[2] Dirk Gevers,[2,3,4] Sarah Preheim,[2] Chiachi Hwang,[5,7] **Matthew Fields**,[5,7] Martin Polz,[2] and **Eric Alm**[2,6,7*] (ejalm@mit.edu)

[1]Computational and Systems Biology Initiative, [2]Department of Civil and Environmental Engineering, [3]Laboratory of Microbiology and [4]Bioinformatics and Evolutionary Genomics Group, Ghent University, Ghent, Belgium; [5]Department of Microbiology, Montana State University, Bozeman, Montana; [6]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts; and [7]Virtual Institute of Microbial Stress and Survival (http://vimss.lbl.gov)

Project Goals: See goals for abstract 36.

To date, modeling the behavior of cellular networks under laboratory conditions has received more attention than modeling how ecological factors affect diversity in natural environments. As we move toward the ultimate goal of integrating laboratory model organism studies with field data, a key challenge will be identifying the geochemical/ecologial factors that underlie community diversity, and the phylogenetic boundaries of natural ecological populations. Thus, computational frameworks for automatically learning models of sequence evolution in the context of metadata (*e.g.*, site geochemistry/ecology) will need to be developed. We present here one such framework: *AdaptML*, a maximum-likelihood-based tool for studying both the sequence evolution and ecological history of a set of gene sequences. To perform this latter task, *AdaptML* employs a hidden-Markov-model-like strategy of assigning gene sequences to unseen states we term "habitats." These habitats are inferred automatically and designed to recapitulate sequence partitioning observed in the wild. AdaptML was initially developed and tested using data from 1027 strains of marine *Vibrio hsp60* gene sequences harvested off the coast of Maine. We show here how *AdaptML* can be used to analyze this dataset and to help build models of *Vibrio* resource partitioning. We have recently applied *AdaptML* to a 16S library collected from a DOE FRC site as part of the ESPP2 project with pre and post biostimulation time points, as a first step toward identifying ecological factors that drive changes in taxonomic diversity and thus organismal fitness in an environment relevant to DOE missions.

62

GTL

VIMSS:ESPP2 Scientific Research Project Management

Amy Shutkin (ashutkin@lbl.gov),[1,2*] **Adam P. Arkin**,[1,2,3] and **Terry C. Hazen**[1,2]

[1]Lawrence Berkeley National Laboratory, Berkeley, California; [2]VIMSS (Virtual Institute of Microbial Stress and Survival) http://vimss.lbl.gov/; and [3]University of California, Berkeley, California

Project Goals: See goals for abstract 36.

The success of a mature scientific research program depends on maintaining an agile Performance Monitoring Project Management structure, which served the flagship VIMSS:ESPP effectively, while implementing more formal risk management oversight practices as proposed new technologies are integrated into the pre-existing core research groups. Project Management offers the stability of a rational and logical process for managing work in a virtual research institute. This (re)new

* Presenting author

research project initiated at three national laboratories and seven universities from coast to coast and is comprised of ~75 individuals collaborating within three core research groups: Applied Environmental Microbiology Core (AEMC), Functional Genomics and Imaging Core (FGIC) and Computational and Systems Biology Core (CSBC). We have found that there is a balance between tight project management to create synergy, focus and continuity to the project and well-tracked individual-investigator-driven initiatives and follow-up that must be struck to maintain creative engagement and productivity of the project team, all while remaining vigilant against scope creep. It has been our experience that frequent and rapid communication at different levels with different media is critical to exploiting the scale and diversity of the team project's capabilities for this multidisciplinary, multi-institutional collaboration.

Included topics in this overview are: Project Management Plan Development and Implementation, Performance Monitoring and Communications, Schedule Development and Execution, Cost Estimation, and Risk Identification and Management.

63

GTL

Systematic Identification of Regulatory Mapping and Optimal Metabolic Engineering Strategies in *Shewanella oneidensis* MR-1

D.K. Byrne[1]* (dbyrne@bu.edu), Q.K. Beg[1]* (qasimbeg@bu.edu), M.E. Driscoll,[1] F.S. Juhn,[1] Y. Shen[2], J.J. Faith,[1] I. Paschalidis,[2] **D. Segre**,[1] and **T.S. Gardner**[1]

[1]Department of Biomedical Engineering and [2]Department of Systems Engineering, Boston University, Boston, Massachusetts

Project Goals: The project goals are to: (1) construct a genome-scale model of transcriptional regulatory pathways controlling metabolic enzymes and stress response systems, (2) complete and validate a metabolic flux model of *Shewanella oneidensis*, and (3) apply a constraints-based analysis to integrated metabolic and regulatory models to optimize electron flux.

Shewanella oneidensis MR-1 is an environmentally ubiquitous, metabolically versatile microbe with a broad capacity for the reduction of metals. These desirable traits have made *Shewanella* a leading candidate for use in microbial fuel cells (MFCs) and environmental remediation of toxic waste. However, practical engineering implementations are limited by our ability to systematically characterize and control key regulatory and metabolic aspects of the organism's physiology.

To this end, we have designed an Affymetrix microarray for *Shewanella* and profiled its genome-wide expression across 255 conditions [1] by varying carbon sources, metallic species, and physiologically relevant factors. We applied our CLR algorithm [2] on this expression compendium to infer the first full-scale regulatory network for *Shewanella*. The resultant network revealed several previously uncharacterized genes as central regulators of the respiratory machinery. *Shewanella* strains containing knock-outs of these genes are currently being investigated to confirm our predictions.

Also, we have developed computational methods to identity optimal strategies to improve *Shewanella* phenotypes related to electrical current production in MFCs. Using a genome-scale metabolic model, developed by the *Shewanella* Federation and led by Jennifer Reed (University of Wisconsin-Madison), we developed and applied algorithms based on linear and nonlinear programming theory to predict optimal nutrients, nutrient compositions, and gene knock-outs that maximize the electrical power output and coulombic efficiencies that *Shewanella* strains produce. Validation of these predictions using batch and continuous bioreactor experiments on wild-type and mutant *Shewanella* strains are currently under way.

The complementary regulatory and metabolic methods developed here will extend current knowledge of the genes, regulators, and metabolites that underlie *Shewanella's* unique respiratory system, and the environmental signals to which they respond. Furthermore, our predictions provide strategies to optimize *Shewanella* for improved metal reduction in real-world applications.

References

1. JJ Faith, ME Driscoll, VA Fusaro, EJ Cosgrove, B Hayete, FS Juhn, SJ Schneider, TS Gardner. (2007) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Research*. [Epub ahead of print].
2. JJ Faith, B Hayete, JT Thaden, I Mogno, J Wierzbowski, G Cottarel, S Kasif, JJ Collins, TS Gardner. (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulatory interactions from a compendium of expression profiles. *PLoS Biology*. 5: e8.

* Presenting author

64

GTL

Towards Genomic Encyclopedia of Carbohydrate Utilization: Case Study in *Shewanella*

Dmitry Rodionov,[1] Chen Yang,[1] Xiaoqing Li,[1] Andrei Osterman[1,2]* (osterman@burnham.org), Olga Zagnitko,[2] Ross Overbeek,[2] Samantha Reed,[3] Margaret Romine,[3] Yanbing Wang,[4] Anna Y. Obraztsova,[4] and **Kenneth H. Nealson**[4] (knealson@usc.edu)

[1]Burnham Institute for Medical Research, La Jolla, California; [2]Fellowship for Interpretation of Genomes, Burr Ridge, Illinois; [3]Pacific Northwest National Laboratory, Richland, Washington; and [4]Department of Earth Sciences, University of Southern California, Los Angeles, California

Project Goals: This project is a component of the *Shewanella* Federation studies aimed at better understanding of the ecophysiology and speciation of this important genus. The main goal of the presented study is to combine genomic reconstruction with physiological experiments to establish and characterize pathways involved in utilization of carbon and energy sources in a collection of *Shewanella* species with completely sequenced genomes.

Carbohydrates are a primary source of carbon and energy for many divergent bacteria. Tremendous variations in carbohydrate pools in different ecological reservoirs are matched by variations within the carbohydrate utilization machinery observed even between closely related species. Our long-range goal is to establish a capability of accurate and comprehensive reconstruction of this machinery from genomic data for hundreds (soon to be thousands) sequenced bacterial genomes. This capability would strongly impact fundamental understanding of microbial ecophysiology and foster applications in bioengineering. Despite the substantial knowledge accumulated in this field, our current ability to project the respective genes and pathways from a handful of model species to others is hampered by the abundance of paralogous protein families of varied specificity, nonorthologous gene replacements and pathway variations. This leads to the accumulation and propagation of imprecise and, often, incorrect genomic annotations in public archives. We address this challenge by combining a subsystems-based approach to genome annotation with experimental validation of selected bioinformatic predictions. The subsystem analysis (as implemented in The SEED genomic platform, http://theseed.uchicago.edu/) is a highly parallel genomic reconstruction of metabolic pathways in hundreds of divergent microbial species supported by the extensive exploration of genome context (conserved operons and regulons). In addition to improving the quality of annotations, this approach allows us to make conjectures about previously uncharacterized genes, pathways and phenotypes. The key inferences are tested by genetic, biochemical and physiological experiments in model species.

We applied this integrated approach to systematically map mono- and disaccharide catabolic pathways in 12 species of *Shewanella* with completely sequenced genomes. This project is a component of the *Shewanella* Federation studies aimed at better understanding of the ecophysiology and speciation of this important genus. We used a set of nearly 250 protein families including transporters, transcriptional regulators and several classes of metabolic enzymes (e.g. sugar kinases, oxidoreductases, epimerases, aldolases, etc) from a collection of carbohydrate utilization subsystems in The SEED to scan for homologs in *Shewanella* genomes. Identified candidate genes were subject of further genome context analysis for their accurate functional assignment and reconstruction of respective pathways. This analysis detected substantial variations in a *sugar diet* between different *Shewanella* species reflecting various aspects of their ecophysiology and evolutionary history. However, more striking are the differences revealed by comparison with a classical model system of *E. coli*. These differences are manifested at various levels, from the presence or absence of certain sugar catabolic pathways to a dramatically different organization of transcription regulatory networks in the central carbon metabolism. The results of this analysis included prediction of several novel variants of carbohydrate utilization pathways (eg for N-acetyglucosamine, sucrose, cellobiose, arabinose and glycerate) and tentative functional assignments for previously uncharacterized gene families (eg within GlcNAc operon SO3503-3507 in *S.oneidensis*, sucrose operon Sfri3988-3991 in *S. frigidimarina* and cellobiose operon Sbal0541-0545 in *S. baltica*). These predictions were verified by phenotype analysis, genetic complementation and biochemical characterization of purified recombinant enzymes. In addition to the specific knowledge of carbohydrate catabolism in the *Shewanella* genus, this study led to a substantial expansion of our current version of the Genomic Encyclopedia of Carbohydrate Utilization. A systematic iterative application of this approach to multiple taxonomic groups of bacteria will further enhance this knowledge base providing an adequate support for efficient analysis of newly sequenced genomes as well as of the emerging metagenomic data.

* Presenting author

65

GTL

Discovery of Novel Machinery for Lactate Utilization by *Shewanella oneidensis* MR-1

Grigoriy Pinchuk,[1] Dmitry Rodionov[2]* (rodionov@burnham.org), Chen Yang,[2] Xiaoqing Li,[2] Samantha Reed,[1] Margaret Romine,[1] Alex Beliaev,[1] and **Andrei Osterman**[2,3] (osterman@burnham.org)

[1]Pacific Northwest National Laboratory, Richland, Washington; [2]Burnham Institute for Medical Research, La Jolla, California; and [3]Fellowship for Interpretation of Genomes, Burr Ridge, Illinois

Project Goals: This project is a component of the *Shewanella* Federation and, as such, contributes to the overall goal of applying the genomic tools to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus.

Lactate is one of the major fermentative metabolism products for many microorganisms and is the most frequently used substrate for experimental studies of respiratory metabolism in *Shewanella oneidensis* MR-1. Consequently, the metabolism of lactate is a key component of the systems-level conceptual model (under development by the *Shewanella* Federation) linking electron transfer networks and central/peripheral carbon metabolism pathways of MR-1. Whereas physiological data has demonstrated a robust growth of *S. oneidensis* on both D- and L- forms of lactate, its genome does not contain orthologs of classical lactate dehydrogenases (LDH) such as D-LDH (gene *dld*) or L-LDH (gene *lldD*) of *E. coli*. We report here the discovery of a novel D- and L-lactate oxidative utilization machinery identified via a comparative genomic reconstruction of *S. oneidensis* MR-1 metabolism combined with physiological, genetic, and biochemical studies.

A hypothetical FeS-containing protein encoded by SO1521 was deemed a candidate for the missing D-LDH based on its presence in the putative operon with an ortholog of lactate permease (SO1522) and its remote homology with the FAD-containing D-LDH from yeast. This prediction was verified by analysis of a SO1521 targeted gene deletion mutant and by genetic complementation of *Escherichia coli* Δ*dld* mutant with a plasmid encoding SO1521. A detailed reconstruction and comparative analysis of *lactate utilization* subsystem including associated operons and regulons, across hundreds of bacterial genomes integrated in The SEED genomic platform (http://theseed.uchicago.edu/FIG/subsys.cgi) led to a conjecture that an adjacent three-gene operon SO1518-SO1520 comprised a previously uncharacterized enzymatic complex for the utilization of L-lactate. Two genes of this operon, SO1519 and SO1520 (previously annotated as a hypothetical FeS oxidoreductase and a ferredoxin-like protein, respectively) appear to form a core of this complex conserved in many divergent bacteria (e.g., uncharacterized operons *ykgEF* in *E. coli* and *yvfVW* in *Bacillus subtilis*). This prediction was validated by assay of targeted gene deletions in *S. oneidensis* and by genetic complementation and testing of the L-LDH enzymatic activity in *E. coli* Δ*lld* mutant overexpressing the SO1518-SO1520 operon. Furthermore, the inability of only two of 19 *Shewanella* sp. with completely sequenced genomes to grow with lactate as sole carbon source are consistent with the results of our comparative genome analysis of these species. These findings, in addition to the identification of previously unknown genes involved in lactate utilization in most *Shewanella* species, broadly impact our knowledge of this important aspect of carbon and energy metabolism in many other bacteria. Additional experiments are in progress to elucidate the details of the novel L-LDH complex in *S. oneidensis* and other species. This project is a component of the *Shewanella* Federation and, as such, contributes to the overall goal of applying the genomic tools to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus.

66

GTL

Multigenome-Based Insights into Respiratory Potential in the Genus *Shewanella*

Margaret Romine[1]* (margie.romine@pnl.gov), Lee Ann McCue,[1] Gretta Serres,[2] and Jim Fredrickson[1]

[1]Pacific Northwest National Laboratory, Richland, Washington and [2]Marine Biological Laboratory, Woods Hole, Massachusetts

Project Goals: This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus.

One of the characteristic features of the Genus *Shewanella* is their remarkable versatility in respiratory metabolism. With the current availability of 17 fully assembled and three partially assembled *Shewanella*

* Presenting author

genome sequences, it is now possible to predict the breadth of substrates that are utilized and to identify common strategies utilized across this genus to achieve their broad diversity in respiratory metabolism. Each type of respiratory system is expected to include cell envelope-localized proteins which contain metal cofactors (e.g. heme, Fe, NiFe, Mo, FeS, Cu) that mediate electron transfer reactions. In order to develop a global overview of respiratory capabilities in *Shewanella*, we are therefore identifying the suite of redox-active proteins present in these strains, and using them as anchoring points for subsequent neighborhood analysis aimed at identifying all of the cellular components necessary to mediate various respiratory electron transfer functions. By cataloging entire subsystems rather than individual components (e.g. terminal reductases) we are able to assess whether novel components are present relative to systems characterized in other bacterial genera and to identify mutations that may be responsible for subsystem inactivation. Results of these analyses are then used to design experimental studies to test phenotypic predictions or protein function. Results from these experiments in turn are used to refine the computational predictions.

Among the proteins deduced from the genome sequences of 20 *Shewanella,* a total of 138 different types of putative *c*-type cytochromes, encoding between one and 12 CXXCH motifs, and 62 different types of other electron transfer proteins (e.g., ferredoxins, hydrogenases, molybdopterin binding oxidoreductases, formate dehydrogenases, flavoproteins) have been identified. Two extreme outliers in redox protein content were found in this group of species: *S. denitrificans* with only 19 predicted *c*-type cytochromes (overall average=40) and 9 other redox proteins (overall average=22) and *S. sediminis* with 83 predicted *c*-type cytochromes and 35 other redox proteins. Besides being able to grow aerobically, the lone shared respiratory capability among all the sequenced *Shewanella* sp. is the reduction of nitrate to nitrite. In contrast to the other *Shewanella* sp., anaerobic respiratory metabolism in *S. denitrificans* appears to be limited to dissimilatory nitrate reduction to N_2O/N_2. Occurrence of the dissimilatory reduction to N_2O subsystem in the sequenced *Shewanella* is rare, with most strains instead encoding genes that mediate nitrate ammonification. Common distinctive themes emerging from our analyses include 1) an expanding number of distinct outer membrane deca- and undecaheme *c*-type cytochromes associated with the metal reducing (mtr) loci, 2) the occurrence of multiple analogous DMSO reductase-like subsystems comprised of a periplasmic decaheme *c*-type cytochrome, an outer membrane localized molybdopterin oxidoreductase, and an outer membrane protein similar to that found in the mtr locus, 3) the occurrence of paralogous respiratory systems, often in immediate proximity to each other, and 4) the dependence of multiple subsystems on the functionality of a single tetraheme quinol reductase (CymA). The large expansion in redox protein content in *S. sediminis* appears to be largely due to the occurrence of paralogous nitrite reductases, DMSO reductase-like systems, and reductive dehalogenases which in part likely reflects its ability to degrade hexahydro-1,3,5-trinitro-1,3,5-triazine (RDX).

67

GTL

Identification and Analysis of Components of the Electron Transport Chains that Lead to Reduction of S-Compounds in *S. oneidensis* MR-1

Eulandria Biddle, Sheetal Shirodkar, and **Daad Saffarini*** (daads@uwm.edu)

Department of Biological Sciences, University of Wisconsin, Milwaukee, Wisconsin

Project Goals: One of the major goals of the funded research is to identify proteins and enzymes that are involved in the reduction of sulfur-containing compounds in *Shewanella oneidensis* MR-1. In addition, the mechanisms that regulate the reduction of sulfur-compounds will be analyzed. A second goal of the project is to proteins that are involved in the regulation of anaerobic respiration in *S. oneidensis*. We have previously identified the cAMP receptor protein as a major regulatory protein of anaerobic respiration. Activation of CRP requires tight regulation of cAMP levels in the cell. One goal of the funded research is to identify phosphodiesterases that may act as oxygen sensors and participate in the regulation of cAMP levels. Additional regulatory proteins that participate in the direct regulation of anaerobic gene expression will also be identified using targeted or transposon mutagenesis. These proteins will be analyzed and their exact functions will be determined.

Shewanella oneidensis MR-1 is able to use S-compounds as electron acceptors for anaerobic respiration. These include tetrathionate, thiosulfate, sulfite, and elemental sulfur. The cAMP receptor protein (CRP) positively regulates the reduction of these compounds in *S. oneidensis*. Two additional regulators are involved in this process. These consist of SO0490 and the two-component system SO4145/SO4147. Mutants that lack SO0490

* Presenting author

exhibit higher levels of sulfur and thiosulfate reduction, but are unable to reduce sulfite. This suggests that this protein plays a dual role in regulating the reduction of S-compounds. SO4144/SO4147 appear to positively regulate the reduction of thiosulfate and sulfur. Interestingly, both SO0490 and SO4145/SO4147 are positively regulated by CRP.

The mechanisms used by *S. oneidensis* MR-1 to reduce S-compounds are not known. It has been suggested that SO4144, an octaheme *c* cytochrome, is a tetrathionate reductase. A mutant that lacks this cytochrome was generated and found to reduce tetrathionate similar to the wild type, indicating that this protein is not the physiological reductase. Mutants deficient in both SO4061 and SO4062 were generated and analyzed. SO4062 is predicted to encode the catalytic subunit of the polysulfide reductase (PsrA), while SO4061 is predicted to encode the Fe-S protein component of the enzyme. Mutants that lack these proteins were deficient in sulfur, thiosulfate, and tetrathionate reduction. These results indicate *S. oneidensis*, unlike other bacteria studied to date, uses the same enzyme to reduce all three compounds.

To identify additional components of the electron transport chain that leads to the reduction of S-compounds, we tested mutants that lack the cytochrome CymA which is involved in electron transfer to Fe(III), nitrate, and fumarate. We also tested mutants that are deficient in cytochrome *c* maturation. Both mutants were able to reduce thiosulfate, sulfur, and tetrathionate similar to the wild type, suggesting that *c* cytochromes are not involved in this process. The polysulfide reductase (PsrABC) responsible for the reduction of sulfur, thiosulfate, and tetrathionate is predicted to be a molybdopterin enzyme that is secreted by the TAT secretion system. This was confirmed by the analysis of mutants deficient in the TAT secretion system and molybdopterin biosynthesis. Both types of mutants were unable to reduce the above-mentioned electron acceptors. Menaquinone deficient mutants were also unable to reduce sulfur, thiosulfate, and tetrathionate. Our results suggest that electrons are transferred directly from menquinones to the terminal reductase, and additional electron transport components do not appear to be required.

In contrast to thiosufate reduction, sulfite reduction appears to require *c* cytochromes. Mutants deficient in cytochrome *c* maturation were deficient in sulfite reduction. Similarly, menaquinone biosynthesis mutants were deficient in sulfite reduction. Analysis of additional mutants is ongoing to identify the sulfite reductase and other components that may be involved in the reduction of sulfite.

"This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus."

68

GTL

Adaptation of *Shewanella oneidensis* MR-1 to its Environment, Insights From Gene Duplication

Margrethe H. Serres[1]* (mserres@mbl.edu), Daniella Wilmot,[1] and Margaret F. Romine[2]

[1]Marine Biological Laboratory, Woods Hole, Massachusetts and [2]Pacific Northwest National Laboratory, Richland, Washington

Project Goals: This project is a component of the *Shewanella* Federation and, as such, contributes to the overall goal of applying the genomic tools to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus.

S. oneidensis MR1 is a model organism for studies on anaerobic respiratory metabolism, with particular emphasis on pathways responsible for the biogeochemical transformation of metals and radionuclides. In addition to the diverse respiratory functions encoded by members of the *Shewanella* genus, they also encode a multitude of genes involved in environmental sensing and regulatory responses. These functions are likely important factors that enable *Shewanella* to thrive in aquatic and sedimentary environments all over the world. Genome sequences of 20 *Shewanella* strains are currently available for study of the ecophysiology and speciation of this genus. The ability of *Shewanella* to grow in a range of environments suggests that its metabolic machinery also has undergone an adaptation or specialization to different environmental niches. We are currently identifying the metabolic capabilities of *S. oneidensis* MR1 from using both a bottom up and top down approach that capitalize on the available genome sequence information and on experimental data generated by members of the *Shewanella* Federation. Our goal is to understand how *Shewanella* sp. have adapted to an aquatic lifestyle by identifying common and distinct types of metabolic pathways present in each of the 20 sequenced strains.

We have built a Pathway/Genome Database, ShewCyc for *S. oneidensis* MR-1 using the Pathway Tools software (1). The database is based on the updated annotation of

* Presenting author

MR1 provided via ongoing efforts of the *Shewanella* Federation. Gene products have been assigned to metabolic pathways and cellular roles according to the MultiFun classification system embedded in the Pathway Tools software. Over 2800 *S. oneidensis* MR-1 proteins, 60% of the gene products, have been assigned to a metabolic pathway or cellular role. Gene products are given multiple assignments when playing more than one role in the cell. ShewCyc also contains genome information generated by members of the *Shewanella* Federation including sequence similar (paralogous) protein groups, protein complexes, transcriptional units, and regulatory interactions.

Paralogous groups of proteins, arisen through gene duplications followed by divergence, encode related functions. Previous analysis of *S. oneidensis* MR-1 included identifying paralogous groups of proteins encoded in the genome (2). Groups with memberships ranging from 2 to 64 were found. Regulators, transporters, transposases, and chemotaxis-related proteins made up the larger groups, reflecting a diverse ability by *S. oneidensis* MR-1 to sense and respond to its environment. Enzymes were found in groups of 14 or less. By analyzing paralogous enzyme groups, namely those that contain members that function in known metabolic pathways, we are trying to identify proteins representing yet undiscovered metabolic capabilities in *S. oneidensis* MR-1. We selected 756 proteins based on their current assignment to metabolic pathways in ShewCyc. Of these proteins, 459 did not have sequence similar matches in *S. oneidensis* MR-1, while 418 belonged to 141 paralogous groups. All proteins in the 141 paralogous groups were included in the dataset, resulting in a total of 858 proteins being analyzed. The size of the metabolism-associated paralogous groups ranged from 2 to 14, with the largest groups encoding SDR-family oxidoreductases, NAD-dependent epimerases/dehydratases, flavoproteins and enoyl-CoA enzymes. We identified 102 proteins in 56 paralogous groups whose functions are related to known enzymes but whose metabolic role remains to be discovered. A further analysis of these proteins and their occurrence in the other *Shewanella* genomes will be presented.

We would like to acknowledge many members of the *Shewanella* Federation for making their data available for inclusion in the ShewCyc database including Lee Ann McCue and Mary S. Lipton (Pacific Northwest National Laboratory), Mike E. Driscoll and Tim S. Gardner (Boston University), Andrei L. Osterman and Dmitri A. Rodionov (Burnham Institute).

References

1. Karp, P., Paley, S., Romero, P. (2002) The Pathway Tools Software. *Bioinformatics* 18:S225-32.
2. Serres, M.H., Riley, M. (2004) Structural domains, protein modules, and sequence similarities enrich our understanding of the *Shewanella oneidensis* MR-1 proteome. OMICS 8:306-21.

69

GTL

Cyclic-di-GMP Signaling in *Shewanella oneidensis* MR-1

Josephine Yu,* Adam Saunders, and **Alfred M. Spormann** (spormann@stanford.edu)

Stanford University, Stanford, California

Project Goals: Cyclic diguanylate (c-di-GMP) is a new bacterial second messenger that controls diverse cellular processes including biofilm formation, motility, and virulence. *Shewanella oneidensis* MR-1 exhibits the unique capacity to transform multiple organic and inorganic electron acceptors in redox-fluctuating environments. Our goal is to elucidate pathways and molecular mechanisms of c-di-GMP signaling in *Shewanella oneidensis* MR-1 involved in sensing and response to changing redox environments.

Cellular level of c-di-GMP is controlled by the interplay of activities of diguanylate cyclases, which were shown to contain amino acid motifs related to the sequence GGDEF, and specific phosphodiesterases, that often carry an EAL or HD-GYP amino acid sequence motif. Current data indicate that effector proteins, which bind c-di-GMP carry PilZ-like domains. The *S. oneidensis* MR-1 genome encodes more than 60 proteins containing GGDEF and EAL domains, including 19 GGDEF or EAL domain proteins that also carry PAS domains, which are predicted to be involved in redox sensing. Using genomic approaches, we have selected several of these that appear to be physiologically significant and have begun to characterize them biochemically and genetically. Interestingly, expression of one PAS/GGDEF/EAL protein at a low level resulted in a motile phenotype, which switches to a sessile phenotype and a loss of red pigmentation at a higher level of induction. *S. oneidensis* MR-1 possesses five proteins that contain a PilZ domain. Deletion of two of the PilZ-encoding genes severely diminished biofilm formation, and enhanced motility in soft agar assays. Interestingly, deletion of another PilZ-related gene gives the opposite phenotype, resulting in enhanced biofilm formation and

* Presenting author

reduced motility, suggesting an antagonistic role. We are currently determining the mechanism through which PilZ domain proteins transduce the c-di-GMP signal using biochemical and genetic methods.

70

GTL

Shewanella baltica: A Model for Examining Specialization Along a Redox Gradient

Jie Deng,* **Jennifer M. Auchtung**,* Jorge L.M. Rodrigues, and James M. Tiedje (tiedjej@msu.edu)

Michigan State University, Center for Microbial Ecology, East Lansing, Michigan

Project Goals: To understand the evolution, ecophysiology, and speciation of different *Shewanella* species.

Shewanella species are thought to play an important role in their ecosystems by coupling the turnover of organic matter with anaerobic respiration of different electron acceptors. Nitrate, an electron acceptor used by many *Shewanella*, is a common contaminant of groundwater as well as several DOE sites (Squillace et al. 2002; Robert and Zachara 1992). One way to remove nitrate from the environment is through the action of denitrifying bacteria. We have begun investigating several strains of *Shewanella baltica* that likely play an important role in the reduction of nitrate. A group of 113 *S. baltica* strains was isolated at different depths from the Baltic Sea, an estuarine ecosystem contaminated with nitrate due to anthropogenic discharges (Ziemke et al. 1998), These *S. baltica* strains represented 77% of the total population of culturable, denitrifying microoorganisms. In addition to the isolation of these organisms, a detailed analysis of the water chemistry and the rates of important environmental processes, such as denitrification, were determined throughout the water column (Brettar and Hofle, 1993; Brettar et al. 2001). These analyses revealed that a relatively stable gradient of terminal electron acceptors (oxygen, nitrate, hydrogen sulfide) was present along the depth of the water column.

A subgroup of 37 *S. baltica* strains was selected for further studies. Phylogenetic analysis of the 16S rRNA gene for all 37 strains indicated that they were members of the same species. Thus, in addition to their potential important role in denitrification, these strains also represent an important dataset for investigating short term evolutionary patterns (i.e., divergence among strains from a single geographic location.) A refined phylogenetic analysis of this group of strains using a less conserved gene marker, namely the B subunit of the topoisomerase (*gyrB*), identified strain specific differences. Strains were grouped into ten statistically supported clades. This result suggests intraspecies genomic heterogeneity. In order to further investigate this possibility and analyze for genome relatedness, we selected seven genes (SO0578, SO0625, SO1771, SO2183, SO2615, SO2706, SO4702) previously characterized as part of the conserved gene core in *Shewanella* for further studies (Konstantinidis et al. 2006). Specific primers were designed and tested against known sequenced *Shewanella* genomes. Multilocus sequence typing of the selected genes has confirmed seven out of ten clades identified by the *gyrB* sequences. How these genomic patterns translate into their ecological physiology is a central question in biology.

In addition to our phylogenetic analyses, we are also investigating differences in the physiology of these 37 strains. Contrary to the way they were isolated, we were not able to detect denitrification from any of the *S. baltica* strains. Although the strains were not capable of denitrification, we suspect that most of the strains will still play an important role in removing nitrate in the environment by reducing nitrate to ammonia, which can then be consumed by other members of the community. *S. oneidensis* MR-1 reduces nitrate to ammonia (Cruz-Garcia et al. 2007) and the genes required for this process are present in the *S. baltica* isolates that have been sequenced. More physiological tests on these strains are ongoing, specifically we will be testing.their growth on different carbon, nitrogen, and phosphorus sources, selection of which will be based on biolog data for two of the *S. baltica* strains. The results of these experiments will highlight physiological differences amongst the strains which, when combined with information about their phylogeny and ecology, should provide insights into the microevolution of this species.

Four *S. baltica* strains were selected for complete genome sequencing based on the following conditions: (1) ecological zone of isolation in the water column (structured as just below the oxic zone with low oxygen and high nitrate (strain OS155), oxic-anoxic transition zone with lower oxygen and high nitrate (strains OS185 and OS223), and the anoxic zone with neither oxygen nor nitrate and low levels of hydrogen sulfide (OS195) and (2) pairwise DNA-DNA reassociation values between 68.2% and 98.2%, representing the same species according to the current bacterial species standards.

Preliminary comparative genomic analysis revealed that although the strains share a large amount of genomic content, the average nucleotide identity between the

* Presenting author

strains ranges from 95.76% (*S. baltica* strains OS155 and OS223) to 97.04% (*S. baltica* strains OS185 and OS195), there are also significant amounts of sequence that vary among the strains. For example, the sequenced genomes of *S. baltica* strains OS155 and OS195 contain 1.1Mb and 1.3Mb of DNA specific to each strain, respectively. These results clearly indicate that these strains are considerably different at the genome level and reveal an unexpectedly large genomic heterogeneity. Experiments are ongoing to determine what effect these differences in genome content have on the cell's physiology and whether they may reflect differential adaptation to microenvironments within the water column. Therefore, through these combined phylogenetic, genomic, and physiologic studies we hope to provide more insight about a major microbial component of a nitrate-contaminated environment while also increasing our understanding of how environment and ecology influence evolution.

References

1. Brettar, I. and M.G. Hofle. 1993. Nitrous oxid producing heterotrophic bacteria from the water column of the central Baltic: abundance and molecular identification. *Mar. Ecol. Prog. Ser.* 94:253-265.
2. Brettar, I., E.R. Moore, and M.G. Hofle. 2001. Phylogeny and Abundance of Novel Denitrifying Bacteria Isolated from the Water Column of the Central Baltic Sea. *Microb. Ecol.* 42:295-305.
3. Cruz-Garcia, C. Murray, A.E., Klappenbach, J.A., Stewart, V., and J.M. Tiedje. 2007. Respiratory Nitrate Ammonification by *Shewanella oneidensis* MR-1. *J. Bacteriol.* 189:656-662.
4. Konstantinidis, K.T., A. Ramette, and J.M. Tiedje. 2007. Toward a more robust assessment of intraspecies diversity using fewer genetic markers. *Appl. Environ. Microbiol.* 72:7286-7293.
5. Riley, R.G. and J.M. Zachara. 1992. Chemical contaminants on DOE lands and selection of contaminant mixtures for subsurface science research. Washington, D.C. U.S. Dept. of Energy, Office of Energy Research. ER-0574T
6. Squillace, P.T., Scott, J.C., Moran, M.J., Nolan, B.T., and D. Koplin. 2002. VOCs, pesticides, nitrate, and their mixtures in groundwater used for drinking water in the United States. *Environ. Sci. Technol.* 36:1923-1930.
7. Ziemke, F., M.G. Hofle, J. Lalucat, and R. Rossello-Mora. 1998. Reclassification of *Shewanella putrefaciens* Owen's genomic group II as *Shewanella baltica* sp. nov. *Int. J. Syst. Bacteriol.* 48 Pt 1:179-86.

71

GTL

Linking *Shewanella* Ecophysiology and Molecular Functions

Anna Y. Obraztsova,[1] **Tatiana V. Karpinets**[2*] (karpinetstv@ornl.gov), Yanbing Wang,[1] Terence B. Kothe,[2] Denise D. Schmoyer,[2] Guruprasad H. Kora,[2] Margrethe H. Serres,[3] Margaret F. Romine,[4] Miriam L. Land,[5] Edward C. Uberbacher,[5] and Kenneth H. Nealson[1]

[1]Department of Earth Sciences, University of Southern California, Los Angeles, California; [2]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [3]Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts; [4]Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington; and [5]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee

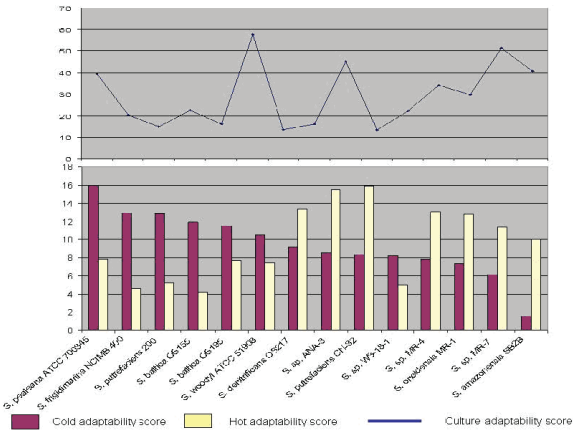
Project Goals: Integrated genome-based studies of *Shewanella* ecophysiology.

Shewanella is an ideal model environmental organism to address a fundamental biological question on linking the adaptive plasticity of the organism to its genotype and to specific cellular processes and molecular mechanisms. Members of the genus inhabit environments that are very different not only in available sources of nutrients and respiration, but also in physical characteristics including temperature, salinity, pH and atmospheric pressure. At present 19 *Shewanella* species from these diverse environments have been sequenced providing an opportunity for comprehensive comparative analysis of their genomes. However, linking this genomic information to complex physiological traits, like bacterial adaptation to culture conditions or to cold and hot temperatures, is challenging, because qualitative characterizations of such complex traits are usually absent.

Objective of the study was to characterize and to relate adaptability of the sequenced *Shewanella* species to different culture conditions to their genomic characteristics. The approach was based on quantifying growth phenotypes and relating them to molecular functions encoded by the genomes. The adaptive plasticity of 15 sequenced *Shewanella* species were characterized experimentally by their growth profiling in temperatures ranging from 0 to 55 degrees C. Bacteria were cultivated aerobically using minimal SF medium with lactate as carbon source. LB medium and marine broth were used only for *S. denitrificans* and *S. woodyi* respectively. The optical density

* Presenting author

(OD_{660}) measurements were used to quantify the adaptive potentials of each species. Specifically, the adaptability to hot temperature was estimated by the Hot Adaptability Score. The score was calculated as the percentage of the accumulated ODs at 30-37 °C if compared with the ODs accumulated at 24 °C, which was optimal for most species. The Cold Adaptability Score calculated as the percentage of the accumulated ODs at 4-12 °C. The Culture Adaptability Score was calculated as the accumulated ODs at optimal growth temperature. For all calculated scores, the greater the score, the greater is the adaptability of the species. This quantification of the analyzed phenotypic traits allowed us to compare the adaptive potentials of the species to culture conditions, cold and hot temperatures, and then to relate these potentials to the enrichment of each genome with different molecular functions, which were characterized in terms of protein families as determined by Pfam. To characterize molecular functions involved in transport and proteolysis in more detail the genomes were also annotated using TransportDB and the peptidase database MEROPS. The number of domains representing each molecular function in each genome was calculated using SQL queries.



Results of the growth profiling (Figure) indicate significant difference in the adaptability of *Shewanella* to temperature and culture not only across species, but also across strains of the same species. The species with a better adaptation to cold temperatures tend to have poorer adaptability to hot temperatures; and the adaptation to culture conditions doesn't correlate to thermal adaptation of the species. These experimental observations indicate that different molecular functions underlie the studied physiological traits. We found that the culture adaptability had the greatest number of molecular functions whose enrichments significantly correlated with the score across species. Most known identified domains represented enzymes involved in proteolysis, including peptidases of different MEROPS families. Another group of the correlated domains revealed the importance of some specific repeats and transposases for successful adaptation of the *Shewanella* genome to culture conditions. Several identified domains represented molecular functions involved in the bacterial type II and III secretion systems and key elements in two-component signal transduction systems. This project is a component of the *Shewanella* Federation and contributes to revealing molecular functions underlying the diverse ecophysiology of this important species.

"The submitted manuscript has been authored by a contractor of the U.S. Government under contract No. DE-AC05-00OR22725. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes."

72

GTL

Motility and Chemotaxis in *Shewanella oneidensis* MR-1

J. Li,[1] W. Song,[1] and **M.J. Ward**[1]* (mjward@jhu.edu)

[1]Department of Geography and Environmental Engineering, Johns Hopkins University, Baltimore, Maryland

Project Goals: Characterize the ecophysiology of *Shewanella oneidensis* by understanding the motility and chemotaxis systems.

Shewanella are a group of microbes that show extreme respiratory versatility. These microbes are also usually found in redox-stratified environments. Taken together, these characteristic features suggest that *Shewanella* ecophysiology is tightly linked to their ability to sense chemical gradients in their environment and respond appropriately. Our role, as members of the *Shewanella* Federation, is to understand how and why these responses occur by analyzing the motility and chemotaxis systems of the model organism, *S. oneidensis* MR-1.

By rotating a single polar flagellum, either clockwise or counter-clockwise, *Shewanella* swim as a series of forward movements and reversals. *S. oneidensis* MR-1 cells have been tracked moving at over 100 μm/sec and with an unstimulated reversal frequency of 0.6 – 0.9 reversals/sec. The flagellar machinery strongly resembles that found in *E. coli*, and 'late' flagellar gene expression is similarly regulated by the alternative sigma factor, sigma 28 (σ^{28}). Using an iterative position specific score matrix-based approach, we have identified several genes that are predicted to have σ^{28} promoters and that are part of a σ^{28}-regulated transcriptional network based on

* Presenting author

transcriptional profiling data from the Shewy Correlation Browser (Gardner Lab, Boston University). The proteins encoded by these genes are not found in *E. coli* and thus may be *Shewanella*-specific components of the motility apparatus.

The chemotactic machinery of *S. oneidensis* MR-1 appears substantially more complex that that of *E. coli*. However, deletion of a single gene, *cheA-3*, results in a non-chemotactic phenotype. Studies involving this non-chemotactic mutant have shown that *S. oneidensis* responds to both electron donors/carbon sources and electron acceptors. Our initial hypothesis, that receptors with redox-sensing PAS domains would dominate behavioral responses by monitoring electron flux through the respiratory electron transport chain, appears to be incorrect, although a ΔSO3404 mutant does show modified aerotactic behavior. Interestingly, another receptor, that we predicted would be required for responses to *N*-acetyl glucosamine, appears to actually be involved in preventing movement towards high concentrations of this electron donor/carbon source. The detailed results of our studies will be presented at the meeting.

We would like to note that conversations with many members of the *Shewanella* Federation have contributed to, or provided direction for, much of our research. We are particularly indebted to Margie Romine (PNNL), Tim Gardner (Boston University), and Igor Zhulin (ORNL).

73

GTL

Integrated Genome-Based Studies of *Shewanella*

Haichun Gao,[1,2] Jingrong Chen,[1]* Yili Liang,[1] Zhili He,[1]* Margaret Romine,[3] Kenneth Nealson,[4] James M. Tiedje,[2] James K. Fredrickson,[3] and **Jizhong Zhou**[1] (jzhou@ou.edu)

[1]University of Oklahoma, Norman, Oklahoma; [2]Michigan State University, East Lansing, Michigan; [3]Pacific Northwest National Laboratory, Richland, Washington; and [4]University of Southern California, Los Angeles, California

Project Goals: The overall goal of this project is to apply the tools of genomics, leveraging the availability of genome sequence for 18 additional strains of *Shewanella*, to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus. To understand these systems we propose to use genome-based approaches to investigate *Shewanella* as a system of integrated networks; first describing key cellular subsystems—those involved in signal transduction, regulation, and metabolism—then building towards understanding the function of whole cells and, eventually, cells within populations.

Shewanella is comprised of more than 40 recognized species inhabiting a wide range of niches planet-wide including the terrestrial subsurface, wetlands, redox interfaces in marine and freshwaters, and cold waters and sediments of the deep sea. Most of these species can utilize multiple electron acceptors including O_2, NO_3^-, S_0, $Mn^{+3,4}$, Fe^{+3}, Cr^{6+}, V^{5+}, DMSO, TMAO, and fumarate. The ability of *Shewanellae* to respire a wide variety of electron acceptors, use the products of other microorganisms and biopolymers such as chitin, and to sense and taxis towards electron acceptors make *Shewanellae* well adapted to living in organic-rich, redox gradient environments. As a part of the *Shewanella* Federation efforts, we have integrated genomic technologies to study various aspects of energy metabolism of two *Shewanella* strains from a systems-level perspective.

Nitrate reduction in *Shewanella oneidensis* MR-1. *S. oneidensis* MR-1 serves as a model for studying anaerobic respiration and electron transport-linked metal reduction. In the genome of *S. oneidensis*, a *napDAGHB* gene cluster encoding periplasmic nitrate reductase (NapA of the NAP system) and accessory proteins and an *nrfA* gene encoding periplasmic nitrite reductase (NrfA of the NRF system) have been identified. However, these two systems appear to be atypical because the genome lacks both *napC* and *nrfH*, which are essential for reduction of nitrate to nitrite and nitrite to ammonium in most bacteria containing these two systems, respectively. In this study, we demonstrated that reduction of nitrate to ammonium in *S. oneidensis* is carried out by these atypical systems in a two-step manner. Unexpectedly, the *napB* mutant exhibited a higher maximum cell density than the wild-type while the *napA* mutant was defective completely in growth on nitrate. Although reduction of nitrate to ammonium in the *napB* mutant is also conducted by NAP and NRF systems, nitrite, the intermediate of the reduction, was not detected through the entire reduction. Further investigation suggests that NapB may be the preferred electron acceptor from a membrane-bound protein which delivers electrons from metaquinol pool to a number of terminal reductases. In an attempt of searching for this membrane-bound protein, both microarrays and mutational analysis have been taken. Results suggest that CymA is likely to be functional replacement of both NapC and NrfH in the nitrate reduction and a novel conceptual model for nitrate reduction is proposed.

* Presenting author

Genomic Array Footprinting in *Shewanella oneidensis* MR-1. Genomic Array Footprinting (GAF) is a high-throughput method to identify conditionally essential genes in microbes by using a combination of random transposon mutagenesis and microarray technology. A GAF was developed for *S. oneidensis* by constructing plasmid pJZ214 from widely used pBSL180. An *S. oneidensis* MR-1 transposon insertion library was generated with pJZ214, containing clones representing ~10^5 independent insertions. This insertion library was subjected to a competitive growth selection in minimal medium with either lactate or glycyl-glutamate as only carbon source for 67 generations in triplicate. Cells were sampled daily. Genomic DNAs from original library and samples were extracted and digested by *Dpn*I to generate DNA fragments of an average size of 400 bp. These fragments were then used as templates for *in vitro* transcription by T7 RNA polymerase and the resulting RNA was labeled by reverse transcription. The labeled cDNAs from samples and original library were cohybridized onto *S. oneidensis* whole-genome microarrays. Meanwhile, conventional microarray was conducted with mRNA from exponential phase cells grown on lactate or glycyl-glutamate. Microarray data from conventional microarray showed that nearly half of genes encoding predicted peptidases (ie. *dcp-1, so1075, so0614*) and a few of genes encoding peptide transporters were up-regulated by the presence of glycyl-glutamate. GAF results from samples with either carbon source revealed a similar number of (~450) genes with reduced signals compared to the original library. These genes belong to a variety of functional categories, suggesting that growth on single carbon source is still a complicated biological process. Although the GAF results were generally consistent with data from conventional microarray analysis, discrepancies were also found. To find the selection pattern of the insertions with respect to the genes' functional classification, systematic analysis is undertaken.

Metal reduction in *S. putrefaciens* W3-18-1. *S. putrefaciens* exhibits an extraordinary ability to reduce irons in various forms. However, little is known about its molecular basis. Predicted counterparts in W3-18-1 to *omcA* and *omcB* of *S. oneidensis* MR-1 are *sputw3181-2445* and *sputw3181-2446*, respectively. Unlike MR-1, W3-18-1 lacks homologs to MtrDEF, which are believed to be composed of a secondary metal reductase. To examine whether Sputw3181-2445 and Sputw3181-2446 of W3-18-1 are functionally equivalent to OmcA and OmcB in MR-1, in-frame *sputw3181-2445* and *sputw3181-2446* individual deletion and double deletion strains were constructed and their metal reduction characteristics were examined. When grown under anaerobic conditions in LB medium containing 20mM sodium lactate as the electron donor and one of following as the electron acceptor: ferric citrate, MnO_2, V_2O_5, MoO_3, Cobalt(III)-EDTA, α-FeOOH, β-FeOOH, Fe_2O_3 or $Fe(OH)_3$, reduction characteristics of these mutants resemble their counterparts of MR-1. These results suggest that OmcA and OmcB in W3-18-1 functions similarly to those in MR-1. Interestingly, the residual metal reduction ability remains even in the *sputw3181-2445* and *sputw3181-2446* double mutant of W3-18-1. A similar observation in MR-1 was presumably due to the presence of MtrDEF. On the basis of that homologs to MtrDEF are missing in W3-18-1 our findings suggest a possible existence of additional unknown metal reductase(s).

74

GTL

Mechanisms of Sulfur Reduction by *Shewanella*

Karlo Lopez, Vinita Lukose, Sneha Vakimudi, and **Edward (E.J.) Crane*** (EJ.Crane@pomona.edu)

Department of Chemistry, Pomona College, Claremont, California

Project Goals: The focus of our group, within the *Shewanella* Federation's overarching goal of understanding *Shewanella*'s ecophysiology, is to determine the pathways of sulfur reduction, and to determine the detailed mechanism of sulfur/polysulfide reduction at the molecular level. Our efforts are divided into 3 main areas of emphasis, which are 1) studies of the native *S. oneidensis* MR-1 sulfur reduction pathway and the native PsrABC complex, 2) construction of recombinant PsrABC complex, produced via heterologous overexpression, for the purpose of detailed mechanistic studies, including site-directed mutagenesis and behavior in immobilized films, and 3) detailed characterization of a recently discovered soluble polysulfide reductase present in *S. lohica* PV-4 and *S. frigidimarina*.

The ability to engage in dissimilatory sulfur reduction (DSR) is a general characteristic of the facultatively anaerobic genus *Shewanella*. While previously thought to be limited to strict anaerobes, the ability to grow while using sulfur as an electron acceptor fits in with the tendency of *Shewanella* to use a wide range of electron acceptors. All of the sequenced genomes of *Shewanella* species contain homologues to the subunits of the polysulfide reductase complex (PsrABC; one *Shewanella* member – *S. denitrificans* – does not contain PsrABC homologues and cannot perform DSR) that has been previously purified from the strictly anaerobic sulfur

* Presenting author

reducer *Wolinella succinogenes*. The basic characteristics of the *W. succinogenes* complex have been deduced in previous studies, however, because of the multiple coenzymes and membrane-bound nature of the complex a detailed understanding of the mechanism of the complex remains to be determined, as do the identities of the direct electron donors, protein or otherwise, to the complex.

The focus of our group, within the *Shewanella* Federation's overarching goal of understanding *Shewanella*'s ecophysiology, is to determine the pathways of sulfur reduction, and to determine the detailed mechanism of sulfur/polysulfide reduction at the molecular level. Our efforts are divided into 3 main areas of emphasis, which are 1) studies of the native *S. oneidensis* MR-1 sulfur reduction pathway and the native PsrABC complex, 2) construction of recombinant PsrABC complex, produced via heterologous overexpression, for the purpose of detailed mechanistic studies, including site-directed mutagenesis and behavior in immobilized films, and 3) detailed characterization of a recently discovered soluble polysulfide reductase present in *S. lohica* PV-4 and *S. frigidimarina*.

The pathway of electrons from donors to their ultimate destination on sulfur is not currently known. Through extensive analysis of mutants, we have shown that in *S. oneidensis* MR-1 DSR is not likely to use any of the multitude of cyctochromes or cytochrome type-proteins (including the "Mtr" proteins) present in the organism. A mutant deficient in menaquinone synthesis, however, does indicate that quinone synthesis is necessary for DSR to occur. We have been able to use membrane extracts to reproduce the sulfur reduction pathway in a cell-free system using both lactate and hydrogen as electron donors. We are currently surveying electron acceptors that will allow us to purify the PsrABC complex (the complex is usually assayed via the reverse reaction, the reduction of compounds by S^{2-}), since further purification of the complex is likely to remove proteins such as hydrogenase that act in the DSR-electron transfer pathway.

In the construction of the recombinant complex, we have overexpressed and purified the subunit that serves as the membrane anchor (psrC), which is purified with the expected bound quinone, suggesting that the soluble overexpressed protein is in an active form, despite its *in vivo* nature as an integral membrane protein. Some of this surprising solubility may be derived from the thioredoxin tag present at the N-terminus of the protein during expression and purification, although the protein remains soluble after the thioredoxin (and poly-histidine) tag is cleaved from the protein. The PsrB subunit has also been overexpressed and purified, and shows UV-visible spectra consistent with the presence of the Fe-S centers observed on the *W. succionogenes* psrB protein. When the PsrC subunit was incubated with a polyanionic film, QCM results indicate that the protein was absorbed into the film, essentially sinking down "into" the film in a manner analogous to the interaction of an integral membrane protein with a membrane. When PsrB was incubated with the PsrC-containing film, the results were consistent with PsrB binding on the surface of the film or the PsrC protein. Both PsrB and C were resistant to being washed off of the film by buffer, and appear to be essentially immobilized within or on the film. We are currently in the process of cloning and overexpressing the PsrA subunit, which contains the molybdopterin site that is likely to be the site of polysulfide reduction. PsrA has been proposed to behave similarly to PsrB in terms of its interaction with PsrC, and the behavior of the PsrB and C subunits (which we had expected to be the most troublesome of the system) inspire confidence that reconstitution of active PsrABC immobilized in a polyanionic film will be feasible, providing a convenient and reproducible system for the in-depth analysis of the redox mechanism of the complex and its molybdopterin center.

During our survey of *Shewanella* genomes we noticed that *S. lohica* PV-4 and *S. frigidimarina* both contained homologues to a family of flavoproteins that our lab had previously studied in the hyperthermophile *Pyrococcus*, although the *Shewanella* protein had an additional "tail" homologous to the polysulfide carrier protein (Sud) from *W. succinogenes.* Results from another laboratory demonstrated that proteins in this family can be involved in DSR (specifically in the reduction of sulfur), and we found that the overexpressed and purified FAD-dependent protein from *S. lohica* PV-4 acts as an NADH-dependent and CoenzymeA-activated polysulfide reductase. This is an entirely new reaction, since the previously observed reaction with the *Pyrococcus* enzyme was elemental sulfur reduction in a CoA-*dependent* manner, with the NADH substrate showing an extremely high K_m (5 mM), while the reaction we've observed is the reduction of *polysulfide* in a CoA-*activated* manner, with a low μmolar K_m for NADH. We are currently in the process of characterizing the mechanism of this enzyme using site-directed mutagenesis and steady-state and presteady-state kinetic techniques.

* Presenting author

75

GTL

Coupled Informatic-Experimental Analyses of Carbon Metabolism Subsystems in *Shewanella*

G. Pinchuk,[1]* O. Geydebrekht,[1] D. Kennedy,[1] S. Reed,[1] M. Romine,[1] C. Yang,[2] D. Rodionov,[2] Andrei Osterman,[2] J. Reed,[3] J. Scott,[4] **A. Beliaev**[1] (alex.beliaev.pnl.gov), and **J. Fredrickson**[1] (jim.fredrickson@pnl.gov)

[1]Pacific Northwest National Laboratory, Richland, Washington; [2]Burnham Institute for Medical Research, La Jolla, California; [3]University of Wisconsin, Madison, Wisconsin; and [4]Dartmouth College-Earth Sciences, Hanover, New Hampshire

Project Goals: *Shewanella oneidensis* MR-1 is a motile, facultative γ-Proteobacterium with remarkable respiratory versatility; it can utilize a range of organic and inorganic compounds as terminal electron acceptors for anaerobic metabolism. More broadly, *Shewanellae* are recognized free-living microorganisms and members of microbial communities involved in the decomposition of organic matter and the cycling of elements in aquatic and sedimentary systems. To function and compete in environments that are subject to spatial and temporal environmental change, *Shewanella* must be able to sense and respond to such changes and therefore require relatively robust sensing and regulation systems. The overall goal of this project is to apply the tools of genomics to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus. To understand these systems we propose to use genome-based approaches to investigate *Shewanella* as a system of integrated networks; first describing key cellular subsystems and then building towards understanding the function of whole cells and, eventually, cells within populations. This information will ultimately be linked to analyses of signal transduction and transcriptional regulatory systems and used to develop a linked model that will contribute to understanding the ecophysiology of *Shewanella* in redox stratified environments.

The genus *Shewanella* is unusually well-adapted to chemically (redox) stratified environments as reflected in the ability to utilize a broad range of electron acceptors via a highly diversified electron transport system. Occupying such niches requires the ability to adapt rapidly to changes in electron donor/acceptor type and availability; hence the ability to compete and thrive in such environments must ultimately be reflected in the organization and flexibility of the central carbon metabolism pathways. Although *Shewanella* species are typically considered to have a relatively restricted substrate range for carbon and energy sources, genome-based analyses revealed multiple pathways for C_{2-6} compounds, amino acids, and fatty acids, reflecting its ecological role as a consumer of organic matter breakdown products in relatively carbon-rich environments that support diverse anaerobic microbial communities. Using controlled cultivation, biochemical, genetic and genomic approaches in conjunction with pathway modeling, we showed that (i) lactate utilization employ previously unknown novel enzymes; (ii) metabolic pathways expressed under different redox conditions utilize pyruvate as a key metabolite and (iii) the pathways involved in ATP production under aerobic and anaerobic conditions fundamentally differ reflecting the amount of energy this organism can generate by oxidative phosphorylation.

Novel enzymes for L- and D-lactate utilization. The first and essential step of lactate utilization is its oxidation to pyruvate, a reaction catalyzed by diverse group of lactate dehydrogenases. The genome of *S. oneidensis* MR-1, the subject of detailed annotation, revealed only one candidate gene, SO0968, predicted to be and NAD^+-dependent lactate dehydrogenase (LDH). Using genetic and biochemical approaches, we found that SO0968 product does not catalyze lactate oxidation but rather encodes a NADH:pyruvate reductase that is involved in pyruvate fermentation and can be used to generate energy in the absence of an electron acceptor. Further research using subsystems-based comparative genomic analysis uncovered novel D- and L-lactate oxidative utilization machinery.

Metabolic pathways expressed under different redox conditions utilize pyruvate as a key metabolite. One of the unresolved problems of *Shewanella* biology is the inability of this organism to couple anaerobic respiration to the oxidation of compounds other than lactate and pyruvate. Growth tests with different organic compounds confirmed that MR-1 can use acetate, succinate, α-oxoglutarate, and propionate as a sole source of carbon and energy under aerobic, but not anaerobic, conditions with fumarate and Fe(III) as electron acceptors. To investigate the underlying mechanisms involved in anaerobic oxidation of carbon substrates, we carried out a comparative analysis of *S. oneidensis* MR-1 metabolism under aerobic, O_2-limited, and anaerobic conditions using integrated genetic, controlled cultivation and functional genomics approaches. Collectively, the results indicate that a shift from aerobic to O_2-limitation leads to fundamental changes in gene expression and enzyme activities that re-wire cells for anaerobic growth.

* Presenting author

Under O_2-limited conditions, pyruvate oxidation was catalyzed exclusively by pyruvate-formate lyase (locus tag: SO_2912), more than 80% of metabolized lactate was converted to acetate through phosphate acetyltransferase and acetate kinase, and biomass yield decreased about 2.5-fold compared to aerobic growth. These changes were accompanied by increased expression of anaerobic respiratory genes including those involved in fumarate, DMSO, and metal reduction. While the full TCA cycle was indispensable for aerobic growth on all compounds tested, the inactivation of α-oxoketoglutarate activity did not affect the anaerobic growth of MR-1 with lactate and fumarate/Fe(III). Moreover, a deletion of the E1 subunit of pyruvate dehydrogenase abolished the ability of MR-1 to grow anaerobically with any of the substrates tested including acetate. It did not, however, impair the ability to grow anaerobically with lactate or pyruvate. In contrast, inactivation of pyruvate-formate lyase did not affect aerobic MR-1 metabolism but impaired anaerobic growth as well as pyruvate fermentation. Overall, our results strongly suggest that oxidation of all tested substrates in *S. oneidensis* MR-1, including acetate, proceeds through pyruvate making this compound a very important oxidative intermediate. The presence of an incomplete TCA cycle and a tendency to oxidize electron donors through pyruvate is likely a key reason why *Shewanella* is unable to use α-ketoglutarate and acetate under anaerobic conditions.

The role of substrate-level phosphorylation under anaerobic and O_2-limited growth of *S. oneidensis* MR-1. One of the fundamental characteristics of *S. oneidensis* metabolism is its inability to use acetate as an electron donor under anaerobic conditions. We extended these previous observations by demonstrating that acetate also cannot be used as an energy source under anaerobic nor O_2-limited conditions. Anaerobic or O_2-limited growth with lactate as the electron donor is accompanied by acetate excretion. We hypothesized that acetate excretion is coupled to ATP production which is catalyzed by acetate kinase (SO_2915). In support of this hypothesis, an MR-1 acetate kinase mutant did not grow anaerobically with either Fe(III)-NTA or fumarate when lactate served as the carbon and energy source. Chemostat experiments also showed that the amount of acetate produced was in inverse proportion to the O_2 flux suggesting that under these conditions, *S. oneidensis* MR-1 growth depends solely on substrate level phosphorylation. We have generated several lines of evidence, including flux balance analysis of *S. oneidensis* metabolism, showing that under conditions of O_2 limitation or fumarate reduction most of ATP is produced from lactate or pyruvate via substrate level phosphorylation (from acetyl phosphate). Our results suggest that, for *S. oneidensis*, the rate of electron transfer to a terminal electron acceptor determines the growth rate and fraction of energy spent on maintenance needs, whereas efficiency of electron transport coupling to phosphorylation partially determines biomass growth yield. Such flexibility of central carbon metabolism would allow *Shewanella* to survive during periods of nutrient-limitation and proliferate rapidly when both electron acceptor(s) and donor(s) are available.

76

GTL

Metabolic Reconstruction of *Shewanella oneidensis*: A Community Resource

Jennifer L. Reed[1]* (reed@engr.wisc.edu), Iman Famili,[2] Sharon J. Wiback,[2] Grigoriy Pinchuk,[3] Margaret R. Romine,[3] Johannes C. Scholten,[3] Joel Klappenbach,[4] and **James K. Fredrickson**[3]

[1]University of California, San Diego, Department of Bioengineering, La Jolla, California; [2]Genomatica, Inc, San Diego, California; [3]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington; and [4]Center for Microbial Ecology, Michigan State University, East Lansing, Michigan

Project Goals: This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus.

Genome-scale network reconstructions account for components and component interactions in biological networks, and are a way in which to collate and analyze data from a variety of sources. Here we report a metabolic reconstruction for *Shewanella oneidensis* MR-1 based on the current genome annotation and primary literature. The reconstruction includes 761 reactions, 789 genes, and 625 metabolites.

The reconstruction was used to build a flux balance model that was used in a variety of computational analyses, including: assessment of growth phenotypes, evaluation of metabolite usage (as substrates or by-products), and prediction of knock-out phenotypes to look at metabolic robustness. The model correctly predicted growth on a variety of carbon and nitrogen sources. In addition, quantitative evaluation of alternative electron acceptors led to the identification of 7 classes of electron acceptors, with differing biomass yields (g D.W. produced per

* Presenting author

mmol electron acceptor consumed). Gene deletion simulations across 10 different environmental conditions with various carbon sources and electron acceptors found that a large fraction of genes were never essential (542 out of 779), while a smaller fraction were always essential (198 out of 779) for growth on these 10 conditions.

Together this work provides a resource that can be used by *Shewanella* researchers and illustrates how reconstructions can serve as a means to evaluate experimental data and generate testable hypotheses to better understand its ecophysiology. This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus.

77

GTL

The *Shewanella* Knowledgebase

Ed Uberbacher[1]* (uberbacherec@ornl.gov), Nagiza Samatova,[1,2] Tatiana Karpinets,[3] Erich Baker,[4] Guru Kora,[3] Andrey Gorin,[1] Mike Leuze,[1] Mustafa Syed,[1] and Denise Schmoyer[1]

[1]Oak Ridge National Laboratory, Oak Ridge Tennessee; [2]North Carolina State University, Raleigh North Carolina; [3]University of Tennessee, Knoxville, Tennessee; and [4]Baylor College of Medicine, Waco, Texas

http://shewanella-knowledgebase.org

The *Shewanella* Knowledgebase is designed to provide a framework for Federation investigators to share, combine and analyze data. The first version of this database was released to the Federation earlier this year. In the last six months or so the effort has focused on a number of issues designed to make the resource more valuable to Federation investigators. These include *(i)* improving the coverage and support for the many data types generated in the SF, *(ii)* improving the data linkage to key investigators and developing procedures to capture their data streams, *(iii)* developing database support for multiple *Shewanella* species and strains, and *(iv)* developing improved linkage to other data resources in the SF, community and to reference materials, *(v)* collecting information on the species available in literature, internet resources and databases, *(vi)* integrating the collected information in the analysis of the experimental data produced by SF.

The project has been successful at defining procedures to keep up with data production at most SF labs. Support for new types of data and projects, including physiology experiments and biolog data have been implemented. These include new capabilities for graphical data associated with experiments, growth curves, and images.

Support for comparative analysis of multiple species has been implemented, including the construction of ShewCyc pathway databases. Tools for species comparison at pathway and genome levels are available and improving. Regulatory data has been integrated from numerous sources as a component of the Knowledgebase. Computational predictions of the regulatory elements in the bacteria were collected from the published literature and from the different Internet resources including Rfam, RibEx, TractorDB, RegTransBase, BioCyc, PromScan, and others. This information was analyzed to identify a set of the basic regulatory classes for their structural presentation in the relational database. They include translated coding sequences, DNA regulator binding sites, sigma factor binding sites, transcription units, promoters, regulons, stimulons, and RNA regulators. The last category encompasses a diverse class of regulators including non-coding and small RNAs, different types of terminators and riboswitches. This component of the Knowledgebase is described in a companion abstract.

Greatly improved linkage to numerous SF and community data sources has been established. A publication mining system that includes a master table of publications with links to the reference sources, authors, and Knowledgebase projects and text search system has also been implemented. Procedures to maintain and update this library are in place. The resource now has *Shewanella* member information, their contributions in terms of data, publications and literature, and a universal resource locator for personal contact.

Web Portal. The *Shewanella* Knowledgebase web portal is a data and knowledge integration environment that allows investigators to query across the *Shewanella* Federation experimental datasets, link to *Shewanella* and other community resources, and visualize the data in a cell systems context. The web portal has many intuitive ways of exploring federation experimental data. The data can be downloaded to the user's computers in the original format. Various data navigation features are also presented to explore the data on the server. *Shewanella* database backend is coupled with powerful system-wide search feature that includes all federation data, publications, literature, etc. The *Shewanella* front-end is built using a combination of Web 2.0 presentation layer technologies. The web portal is built with HTML 4.0, CSS & Script.aculo.us javascript library. The content is mostly

* Presenting author

dynamically generated using Java Server Pages Standard Tag Library (JSTL).

Data Analysis. The user interface has many intuitive guides and wizards to explore *Shewanella* experimental results. It features many experimental data comparative analysis modules that perform one-on-one analysis with diverse sets of biological data, with corresponding visualization capabilities at various data aggregation levels and in different biological contexts. It also provides a unified set of integration analysis tools that currently support ShewCyc pathways and pathway group categories. Future releases will include KEGG pathways, TIGR roles, and GO ontologies for exploring data.

Data Visualization. Various data visualization schemes are provided to display the results of *Shewanella* Federation experiments. One of the viewers compares relative expression data at the gene level, while other viewers compares the average or percentages of the under/over expressed genes in a pathway or pathway group. These viewers are also cross-referenced to Pathway Tools software which contains reference pathways for multiple *Shewanella* strains.

Computing Infrastructure. The computing infrastructure for the database has been further developed. The *Shewanella* web portal has a multi-tier architecture with each tier hosted on a separate server and designed to work independently. It is build with open source MySQL database and is hosted on a power/data redundant server at ORNL. The *Shewanella* backend database provides numerous security features like row level locking, user level access privileges etc.

78

GTL

Comprehensive Integration of Regulatory Data in the *Shewanella* Knowledgebase

Mustafa H. Syed,[1] **Tatiana V. Karpinets**,[2] Denise D. Schmoyer,[2] Michael R. Leuze,[2] Nagiza Samatova,[2,3] and Edward C. Uberbacher[1]* (uberbacherec@ornl.gov)

[1]Biosciences Division and [2]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee and [3]North Carolina State University, Raleigh, North Carolina

Project Goals: Data integration and sharing for the *Shewanella* Federation.

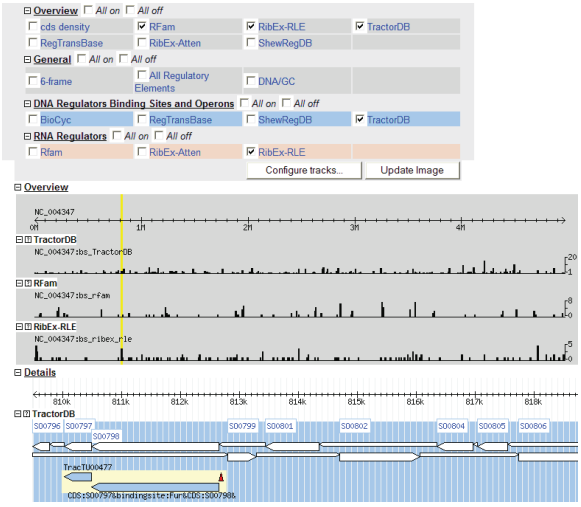
It has become increasingly evident over the recent years that cell regulation is a complex multilayer process encompassing all steps of the information processing in the cell and including a diverse set of regulatory elements. Untranslated regions in *Shewanella* may be considered as a complex regulatory system that can both produce and integrate the regulatory information in the cell, thereby adjusting cellular processes to the environment. Experimental identification of the regulatory elements, however, is very laborious, and as a result numerous computational methods have been developed to predict elements. even in the model organisms it involves a lot of computational predictions. The results of many such predictions for *Shewanella* are found in many different regulation-related databases. Thus, highly relevant regulatory information related to *Shewanella* is scattered across publications and the different Internet resources. At present there is no accepted infrastructure that allows one to integrate, visualize, and analyze the regulatory information based on different computational predictions for a particular organism. This infrastructure, however, is very important to gain insight into *Shewanella* regulation at the level of each individual gene, to check the predictions and to integrate the diverse regulatory information in the analysis of experimental data, especially those obtained by large-scale technologies, like microarray or proteomics.

Objective of this work was to develop an infrastructure for collection, storage, and visualization of the regulatory information relevant to *Shewanella oneidensis* MR-1 and to integrate this information with numerouse experimental datasets submitted to the *Shewanella* Knowledgebase (http://shewanella-knowledgebase.org). The computational predictions of the regulatory elements were collected from the published literature and from

* Presenting author

the different Internet resources including Rfam, RibEx, TractorDB, RegTransBase, BioCyc, PromScan, and others. This information was analyzed to identify a set of the basic regulatory classes and define their representation in the Knowledgebase. Elements include translated coding sequences, DNA regulator binding sites, sigma factor binding sites, transcription units, promoters, regulons, stimulons, and RNA regulators. The last category encompasses a diverse class of regulators including noncoding and small RNAs, different types of terminators and riboswitches.

Visualization of the collected *Shewanella* regulatory information was implemented using Gbrowse from the Generic Model Organism Database Toolkit http://iubio.bio.indiana.edu /gmod /gbrowse/), which was configured for the *Shewanella oneidensis* MR-1 genome and adjusted to a specificity of the collected information (see Figure). Options currently available in the regulatory element browser include (a) an overview of all regulatory elements in the genome with scrolling to any selected region, (b) presentation of each type of element on different tracks or all elements together on one track, (c) different types of zooming, d) simultaneous bird's eye and detailed views of the genome, (e) brief information on each element by popup balloons and (e) detailed information on each regulatory element including decorated (colored conserved regions) FASTA sequences. Users can download regulatory sequences and feature tables in various formats.



79

GTL

Adaptive Evolution of *Geobacter* to Optimize Electricity Production and Bioremediation and to Elucidate Complex Physiological and Ecological Phenomena

Zara Summers, Kelly P. Nevin* (knevin@microbio.umass.edu), Shelley A. Haveman, Hana Yi, Trevor L. Woodard, and **Derek R. Lovley** (dlovley@microbio.umass.edu)

University of Massachusetts, Amherst, Massachusetts

Project Goals (Abstracts 79, 84-86): The goal of this project is to provide computational tools to predictively model the behavior of two microbial communities of direct relevance to Department of Energy interests: 1) the microbial community responsible for in situ bioremediation of uranium in contaminated subsurface environments; and 2) the microbial community capable of harvesting electricity from waste organic matter and renewable biomass. Based on our current rate of progress, it is expected that within five years it will be possible to predict the in situ growth and activity of the relevant microorganisms in the environments of interest solely from relevant geochemical data and to describe in detail the in situ metabolic state of the microorganisms from environmental gene expression data. Furthermore, these computational tools will be able to predict the response of the microbial community to environmental manipulations or manipulation of the genome of the relevant organisms, allowing rational optimization of in situ uranium bioremediation or electricity harvesting via environmental or genetic engineering.

Geobacter species are the most effective microorganisms known for converting organic compounds to electricity in microbial fuel cells and are important in the bioremediation of subsurface environments contaminated with organic and/or metal contaminants. However, the optimal conditions for power production and contaminant bioremediation are significantly different from the low-nutrient, slow growth conditions under which *Geobacter* species are likely to have evolved. Furthermore, genome-scale *in silico* modeling of the metabolism of *Geobacter* species has suggested that *Geobacter* species have metabolic potential far beyond that previously demonstrated in culture. Therefore, it was hypothesized that by exerting the appropriate selective pressure on *Geobacter* species it might possible to increase their effectiveness for practical

* Presenting author

applications as well as develop a better understanding of complex physiological processes within these organisms.

For example, although the culture-collection strain of *Geobacter sulfurreducens* produces current at higher power densities than any other microorganism known, the effectiveness of power production was significantly improved with continuous selective pressure for rapid growth on electrode surfaces. In a relatively short time it has been possible to increase overall current production in microbial fuel cells 600%, but even more remarkably, this increased current is produced by far fewer cells. Whereas the culture-collection strain requires a thick biofilm to produce maximal current, similar biofilms are not seen with the evolved strain. Thus, the current production per mg of cell protein is at least 30-fold higher than with the culture-collection strain and continues to improve. The genomes of these evolved strains are currently being sequenced to determine what beneficial mutations have contributed to this remarkable increase in current production capacity.

Another *Geobacter* strain was evolved to transfer electrons to fuel cell anodes at potentials significantly lower than those of the culture-collection strain. This strain has deleted over 6% of its genome. It appears to be adapted not only for electron transfer at low potential, but also for more rapid growth on electrodes and Fe(III) oxide. Potential additional physiological consequences of such a large gene loss are currently under investigation.

More is known about the evolution of another strain of *G. sulfurreducens* that was adapted for rapid extracellular electron transfer to Fe(III) oxide and grows 10-fold faster on Fe(III) oxides than the culture-collection strain. Five mutations were identified in the adapted strain. These were in regulatory genes and genes encoding proteins for fumarate transport and outer membrane protein biogenesis. Microarray analysis comparing gene transcript levels in the adapted strain and the wild type suggested an upregulation in expression of genes encoding: the electrically conductive pili; other extracellular electron transport proteins; and TCA-cycle enzymes. Gene knock-out studies suggested that the pathway for extracellular electron transfer is somewhat different in the evolved strain and biochemical studies demonstrated that the adapted strain has a greater abundance of loosely bound, outer-surface *c*-type cytochromes.

A metabolic feature of *Geobacter* species limiting current production from complex wastes, and restricting the substrates that can be used to promote *in situ* uranium bioremediation, is the limited range of substrates that these organisms can metabolize. For example, the culture-collection strain of *G. sulfurreducens* is unable to utilize energy/electron-dense compounds, such as sugars and glycerol. Furthermore, although the culture-collection strain of *G. sulfurreducens* can grow on lactate, its doubling time with lactate as the electron donor and fumarate as the electron acceptor is 24 h whereas the doubling time predicted by the *in silco* model for this growth condition is 2.6 h. Repeated transfer of *G. sulfurreducens* in lactate-fumarate medium for over 500 generations reduced the doubling time to 5 h. Furthermore, a lactate-adapted strain was recovered that not only had improved growth on lactate, but also could utilize pyruvate, as well as a number of precursors to pyruvate, such as a variety of sugars and glycerol. The evolution of a strain of *G. sulfurreducens* that can utilize electron-dense fuels, such as sugars and glycerol, that are important components of biomass or wastes (glycerol is a major waste product of biodiesel production) has important implications for current production and bioremediation. Furthermore, it raises an important physiological/ecological question of why *Geobacter* species do not utilize these compounds in soils and sediments, but rather rely on fermentative microorganisms to convert these substrates to acetate and other fermentation products that *Geobacter* species then utilize.

When *G. sulfurreducens* was grown in continuous culture for a year with Fe(III) citrate as the electron acceptor, two of the four strains examined lost the capacity for fumarate reduction whereas strains that had been grown in continuous culture for a year with fumarate as the electron acceptor retained the capacity for Fe(III) reduction. The loss in the capacity for growth on fumarate was associated with mutations in a sigma 54-dependent DNA-binding response regulator that is located immediately upstream of the gene for the fumarate transporter. The loss of a functional fumarate transporter is one of the responses that was also observed in the strain mentioned above which was adapted for rapid growth on Fe(III) oxide. These results suggest that a functional fumarate transporter is deleterious for growth with Fe(III) as the sole electron acceptor. The fact that the fumarate transporter is found in fresh *Geobacter* isolates from a diversity of sedimentary environments suggests that, in nature, a fumarate transporter is a beneficial feature. Further investigation of this hypothesis is underway.

The relative ease in developing strains for enhanced extracellular electron transfer and range of substrate utilization with adaptive evolution contrasts with our previous inability to genetically engineer such large improvements in respiration and metabolism via the enhanced expression of what was considered to be the necessary genes. These results demonstrate that for complex, highly

* Presenting author

regulated, poorly understood, physiological properties, adaptive evolution may be the superior initial design tool. Once the mutations leading to the desired modifications are elucidated via adaptive evolution studies, leading to a better understanding of the physiology, then engineering via other strategies may be possible. These studies also suggest that adaptive evolution studies can provide new insights into the physiology and ecology of *Geobacter* species living in subsurface environments.

80

GTL

Bioinformatic Analysis of Transcription Regulation of *Geobacter sulfurreducens*

Julia Krushkal[1]* (jkrushka@utmem.edu), Jose F. Barbe,[1] Ching Leang,[2] Toshiyuki Ueki,[2] Katy Juarez-Lopez,[3] Jeanette Peeples,[1] Yanhua Qu,[1] Ronald M. Adkins,[1] and **Derek R. Lovley**[2] (dlovley@microbio.umass.edu)

[1]University of Tennessee Health Science Center, Memphis, Tennessee; [2]University of Massachusetts, Amherst, Massachusetts; and [3]Instituto de Biotecnología/UNAM, Cuernavaca, México

Project Goals: The overall purpose of this project is to develop experimental and computational tools to predictively model the behavior of complex microbial communities involved in microbial processes of interest to the Department of Energy. The five year goal is to deliver in silico models that can predict the behavior of two microbial communities of direct relevance to Department of Energy interests: 1) the microbial community responsible for in situ bioremediation of uranium in contaminated subsurface environments; and 2) the microbial community capable of harvesting electricity from waste organic matter and renewable biomass. The research in this abstract summarizes research under Subproject IV. The purpose of this subtask is to use computational methods to better understand transcriptional regulation of the expression of environmentally relevant genes in *Geobacter* species.

Geobacter sulfurreducens is a pure culture model representative of the family *Geobacteraceae,* a group of important organisms that participate in bioenergy production and in environmental bioremediation. Our research employs bioinformatic tools to discover transcriptional regulatory mechanisms that are important for *Geobacter* physiology, its response to environmental changes, and its role in bioenergy generation.

Our recent studies have focused on the *G. sulfurreducens* RpoN (σ^{54}) regulon. RpoN is an essential sigma factor, a subunit of RNA polymerase, involved in a variety of cellular processes in *G. sulfurreducens.* Using computational tools, we identified sequence sites in the *G. sulfurreducens* genome that likely represent promoter elements recognized by RpoN. These promoters, predicted using sequence information, were ranked according to their scores, their location, and orientation relative to their target operons, and the change in expression of their target operons in the strain of *G. sulfurreducens* overexpressing the *rpoN* gene as compared to the wild type. We identified RpoN-regulated promoters in the upstream regions of both significantly up- and downregulated operons, which suggests competition by RpoN with other sigma factors. A number of RpoN-regulated promoters with high scores were found upstream of upregulated operons encoding ABC-type branched-chain transporter system proteins and ribosomal proteins, and also upstream of downregulated operons related to flagellar biosynthesis, nitrogen regulation, and ion transport. RpoN-regulated promoters were also identified upstream of operons encoding c-type cytochromes, DNA polymerase subunits, signal transduction components, and other proteins. We further investigated the function of RpoN by analyzing the presence of RpoN-regulated promoter elements near genes with significant changes in expression levels in several deletion mutants of *G. sulfurreducens* (e.g., *omcB* mutant adapted to growth on soluble Fe(III), *pilR* mutant, and others). Several predicted RpoN-regulated promoters have now been experimentally validated. Identification of RpoN-regulated promoters in the genome, combined with functional studies of the RpoN regulon carried out by other members of the *Geobacter* Project, allows a deeper understanding of the importance of RpoN as a global transcriptional regulator of *G. sulfurreducens.*

Corollary to our studies of the RpoN regulon, we are also investigating the PilR regulon. PilR is an enhancer binding protein, which acts cooperatively with RpoN in transcriptional regulation. PilR is involved in the expression of the *pilA* gene, whose product is a structural component of pili, also referred to as nanowires, which are electrically conductive and are required for Fe(III) oxide reduction as well as optimal current production in microbial fuel cells. Using sequence information and microarray analysis of PilR deletion mutant, we used bioinformatic analyses to predict multiple PilR-regulated sites that may affect transcription of operons related to biosynthesis, assembly, and function of pili and flagella, and to cell wall biogenesis. We also documented the co-occurrence of predicted PilR regulatory sites with RpoN-regulated promoters. Experimental validation of these computational predic-

* Presenting author

tions is now underway. This computational discovery of PilR-regulated sites, if confirmed experimentally, along with our knowledge about RpoN-regulated promoter elements will aid in the understanding of the intrinsic nature of molecular mechanisms regulating production and assembly of pili and flagella that are important for Fe(III) reduction and the production of electric current.

We are also continuing our effort to catalogue predicted and reported transcriptional regulatory interactions in *Geobacter*, and we are developing and applying sophisticated software and data mining tools to better analyze and report these interactions. For this purpose, we are continuing to develop and update a database and an accompanying online query system, GSEL (*Geobacter* Sequence Elements), created by our group. This database compiles information on transcription regulatory elements in the genome of *G. sulfurreducens*. We are currently expanding the GSEL database in order to allow its timely update of operon predictions, to incorporate new information on predicted or experimentally validated genome regulatory sites and interactions, and to allow manual curation of database entries, as well as future addition of new search capabilities. The development of the GSEL database and online query system is currently performed using SQL-based database management systems, PHP technologies, and Python frameworks. In addition, we are continuing our ongoing role in the prediction of operon organization of newly sequenced genomes by the *Geobacter* Project. These predictions are being used to better understand transcriptional regulatory mechanisms of gene co-regulation in *Geobacteraceae*. The new information about operon organization is being incorporated in the GSEL database and is also being extensively utilized in experimental studies of *Geobacteraceae*.

81

GTL

Computational Modeling of Central Metabolism of *Geobacter* and Related Species

Jun Sun,[1]* Kai Zhuang,[2] Karthik Srinivasan,[2] R. Mahadevan,[2] Olivia Bui,[1] Tom Fahland,[1] Iman Famili,[1] Mounir Izallalen,[3] Carla Risso,[3] Shelley Haveman,[3] Christophe Schilling,[1] and **Derek Lovley**[3] (dlovley@microbio.umass.edu)

[1]Genomatica, Inc., San Diego, California; [2]University of Toronto, Toronto, Ontario; and [3]University of Massachusetts, Amherst, Massachusetts

Project Goals: The goal of this project is to incorporate the wealth of biological data and findings into in silico models, which can describe and predict the likely behavior of *Geobacter* species under various environmental conditions during in situ uranium bioremediation or on the surface of energy-harvesting electrodes. We have implemented an iterative modeling and experimental program to develop a model for central metabolism in *Geobacter sulfurreducens* which accurately predicts the growth and metabolism in this organism under various environmental conditions and has provided a quantitative analysis of factors controlling the competition between Fe(III) reducers in subsurface environments as well as aiding significantly in the functional analysis of the *G. sulfurreducens* genome. In these studies the genomes of various *Geobacteraceae* available in pure culture or assembled from environmental genomic data will be compared on a functional basis by determining the likely metabolic capabilities and responses of these organisms under different environmental conditions with constraint-based in silico modeling. Further, the general methods and approaches developed in these studies will be used to develop genome-based models of other microbial communities that are relevant to processes in the subsurface sediments or on the surface of energy-harvesting electrodes.

Computational modeling on *Geobacter* and related species were conducted in order to better understand the physiology and ecology of *Geobacteraceae* involved in bioremediation and electricity generation from waste organic matter and renewable biomass.

Geobacter metallireducens is of interest because it represents a model for the *Geobacter* species that carry out bioremediation of organic and metal contaminants in subsurface environments. The genome-scale metabolic

* Presenting author

model of *G. metallireducens* was further refined to include 747 genes and 697 reactions. Detailed examination of the refined *G. metallireducens* model suggested that its central metabolism contains several potential ATP-consuming futile cycles, involving energy-inefficient reactions that are not present in the *G. sulfurreducens* model. These *G. metallireducens* unique reactions include the acetyl-CoA synthetase, acetyl-CoA hydrolase, and phospho-enolpyruvate carboxylase reactions. We hypothesized that the energy-inefficient reactions and futile cycles might function to balance key metabolite pools to allow higher fluxes through the TCA cycle for energy production. Experimental biomass yield of *G. metallireducens* growing with pyruvate was lower than optimal *in silico* biomass yield but matched simulation results when fluxes through the futile cycles were assumed. Microarray data of *G. metallireducens* growing with benzoate and acetate indicated that genes coding these unique reactions were up-regulated by benzoate. These results suggested that the futile cycles were likely turned off during *G. metallireducens* growth with acetate for optimal energy utilization, but were up-regulated during growth with complex electron donors to improve flux through TCA cycle for fast energy generation.

Growth of *G. metallireducens* with different electron donors and electron acceptors were simulated using the *in silico* model. Simulation results indicated that: 1) under donor limiting conditions, aromatics compounds such as toluene and cresol allowed the highest biomass yield per substrate and acetate allowed the least; 2) under acceptor limiting conditions, pyruvate produced the highest biomass yield per acceptor; 3) complete oxidization of aromatic compounds required much more electron acceptor than pyruvate; 4) nitrate as electron acceptor resulted in higher biomass yield per substrate or per electron acceptor than Fe(III) or fumarate (in a *G. metallireducens* strain with a dicarboxylic acid transporter to allow the utilization of fumarate). These simulations provided a fast and cost-effective way to understand the metabolism of *G. metallireducens*.

One of the key results from continued genome-scale modeling of *G. sulfurreducens* was the impact of the global proton balance on the biomass yields during reduction of extracellular electron acceptors. In contrast to fumarate reduction, protons generated during acetate oxidation are not consumed during the reduction of Fe(III) as the terminal electron acceptor. The energetic cost of pumping these protons to maintain the membrane potential was implicated in lowered biomass yields during growth with extracellular electron acceptors. We have experimentally evaluated the rate of proton exchange during growth of *G. sulfurreducens* to confirm the model prediction of proton generation rate.

Furthermore, constraint-based modeling was applied to the investigation of metabolism of *Pelobacter carbinolicus* in the *Geobacteraceae* family. The reconstructed *P. carbinolicus* model contains 740 genes and 705 reactions, and shared 539 reactions with *Geobacter sulfurreducens* metabolic model. The unique reactions of *P. carbinolicus* model reflected some unique metabolic capabilities such as fermentation growth, a second pathway for proline biosynthesis, and the additional reactions in purine biosynthesis. Microarray data were utilized to understand the redundancy associated with some reactions and pathways. To validate the reconstructed model, *P. carbinolicus* model was tested by simulating published growth conditions including fermentations, syntrophic growth, and Fe(III) reduction. Simulation results matched well with experimental data and indicated the accuracy of the current model.

Rhodoferax ferrireducens is a dissimilatory metal-reducing microorganism which has recently been shown to be abundant in some uranium-contaminated subsurface environments and also has the unique abilitiy to effectively convert sugars to electricity. The reconstructed *R. ferrireducens* genome-scale model contains 737 genes and 756 reactions. The *R. ferrireducens* metabolic model can grow with various electron donors, including acetate, pyruvate, lactate, malate, glucose, benzoate, etc. To understand *R. ferrireducens* growth with these different electron donors, *in silico* simulations were performed with the *R. ferrireducens* model. Under both electron donor and acceptor limiting conditions, glucose allowed the highest biomass yield per electron donor or per electron acceptor. The *R. ferrireducens* modeling research also helped to understand the experimental results that *R. ferrireducens* could not grow with only glucose, but could grow with fumarate serving as both electron donor and acceptor. *R. ferrireducens* appears to be missing a reversible acetaldehyde dehydrogenase that prevents it from growing via glucose fermentation. However, *R. ferrireducens* contains all the genes required to utilize fumarate as electron acceptor allowing *R. ferrireducens* to grow via fumarate fermentation.

In order to better understand the ecology and function of dissimilatory metal-reducing communities in contaminated subsurface environments the dynamics of competition in a metal reducing tri-culture of *R. ferrireducens*, *G. sulfurreducens*, and *E. coli* was modeled. *R. ferrireducens* can oxidize both glucose and acetate while *G. sulfurreducens* is dependent on fermentative microorganisms to produce acetate from glucose. Four sets of two-thousand

* Presenting author

simulations with randomized parameters were performed. The randomized sets are defined by having either high or low maximum glucose feed rate. For the sensitivity analysis, kinetic parameters were varied between 25% and 400% of their literature values. *G. sulfurreducens* dominated in the majority of simulations, confirming the typical observation in subsurface environments. Under glucose-rich conditions, glucose feed concentration shows the highest influence on the *Geobacter* to *Rhodoferax* ratio (G/R ratio)—more glucose allows fermenters to exploit the r-strategy more efficiently. However, if the glucose input flux is low, the G/R ratio is mostly dependent on the initial concentration of *G. sulfurreducens*. This result suggests that under these conditions, the fermenters' r-strategy fails and *G. sulfurreducens* success becomes highly dependent on its own initial concentration.

It is surprising that *R. ferrireducens* are outnumbered by *G. sulfurreducens* in anaerobic aquifers and sediments, even though *R. ferrireducens* is metabolically more versatile than *G. sulfurreducens*. Based on simulation results, we propose that the fermenters and *G. sulfurreducens* out-compete *R. ferrireducens* in utilizing glucose and acetate. Under substrate rich conditions, fermenters first out-compete *R. ferrireducens* for glucose using r-strategy, and convert glucose to acetate; then, *G. sulfurreducens* out-competes *R. ferrireducens* for acetate using r-strategy. Under conditions with limited substrate addition, the co-culture lacks sufficient substrate to exploit the r-strategy efficiently, leaving a niche for *R. ferrireducens*.

82

GTL

Continued Identification of Small Non-Coding RNAs and Acceptance Rate Studies in Members of the *Geobacteraceae*

Barbara Methé[1]* (bmethe@jcvi.org), Robert Deboy,[1] Sean Daugherty,[1] Manolito Torralba,[1] Kelly Nevin,[2] Jonathan Badger,[1] and **Derek Lovley**[2]

[1]The J. Craig Venter Institute, Rockville, Maryland and [2]University of Massachusetts, Amherst, Massachusetts

Project Goals: The *Geobacter* project seeks to apply a systems level biological approach to the study of members of the *Geobacteraceae*. Specific goals of the project presented in this abstract are to apply comparative genomic analyses to understand important aspects of regulation and evolution of these organisms.

Members of the *Geobacteraceae* continue to be the subject of comprehensive studies related to their ability to degrade carbon compounds including many contaminants with the reduction of Fe(III). Further interest in these organisms stems from the practical biotechnological roles that they can play as agents of bioremediation and in energy production. We have begun several analyses that leverage information available from the genome sequences of multiple members of the *Geobacteraceae* to obtain new insights into their evolution and regulation. An update of these analyses is provided here.

In order to better understand the regulatory network of the model organism, *G. sulfurreducens*, as well as other members of the *Geobacteraceae*, we have initiated a study of chromosomally located small non-coding (sRNAs). Increasing evidence suggests that sRNAs exist in numerous organisms where they play important regulatory roles including responses and adaptations to different stresses. We have extended our predictions of sRNA candidates through continued use of the program, sRNAPredict, along with custom written scripts, to compare the *G. sulfurreducens* genome to a total of seven members of the *Geobacteraceae* (*G. metallireducens*, *G. uraniumreducens* *G. bermidjiensis*, *G.* sp. FRC-32, *G. lovleyi*, *P. carbinolicus*, and *P. propionicus*). The sRNAPredict program utilizes BLAST files generated from intergenic regions (IGRs) of *G. sulfurreducens* against the chromosomes of the other test strains. Positive IGRs were cross-referenced with files generated from TransTerm, RNA motif and QRNA.

From the updated searches, 88 predictions were generated for *G. sulfurreducens*: None of the predictions from *G. sulfurreducens* were found in all of the other seven genomes tested although 17 were found in six of the seven. However, it cannot be ruled out that the distributions of sRNAs could change as three of the test genomes used in these searches are still in a draft phase. The majority of the predictions (34) were found in only one of the other seven genomes. Distributions for the remaining predictions are given as follows: five genomes (4), four genomes (7), three genomes (13), and two genomes (13). Of the 88 predictions, seven have matches to covariance models to non-coding RNAs in the Rfam database. This includes a match to a 6S RNA which is believed to associate with the RNA polymerase holoenzyme containing the sigma70 factor and repress expression from a sigma70-dependent promoter. The presence of this sRNA has been experimentally verified through the use of Northern blot hybridizations and sequence specific primers designed to amplify the sRNA from a population of cDNAs. Although the 6S RNA has been characterized as important in stationary phase growth

* Presenting author

we have seen evidence of its expression in both log and stationary phase growth in *G. sulfurreducens*.

An important question yet to be answered comprehensively is whether or not the physical location of a particular family of sRNAs in a query genome is a random event when compared to a family of genomes, or is a "syntenic" relationship maintained. This question has relevant implications in terms of genome evolution and in enhancing the identification of sRNA predictions. An examination of this question was initiated by evaluating whether or not a predicted sRNA in *G. sulfurreducens* and its homolog in a test genome maintain neighboring genes. Eight predictions were analyzed that matched at least three test genomes. Five were found to be associated with a particular gene or set of genes, and each of these sRNA predictions maintained their relative orientation to the neighboring gene(s) in *G. sulfurreducens* and the test genomes. The predictions which "failed" were short sequences with blast e-values above 1e-5 in some test genomes. These might be pursued by adjusting cutoffs and looking at neighboring genes to confirm correct annotation in the test genomes. These results suggest that sRNAs of a particular family may be associated with a similar gene or gene(s) in a family of genomes.

Evolutionary processes are important forces that drive changes in microbial genomes and they provide a lens through which to better view and understand them. Since advantageous mutations are fixed in a population more rapidly than neutral mutations the rate of nonsynonymous (amino acid replacement) substitution will exceed that of synonymous (silent) substitution if advantageous selection plays a role in the evolution of the protein in question. One method to detect positive selection is to determine if the number of substitutions per nonsynonymous site is significantly greater than the number of substitutions per synonymous site.

To examine sites within predicted proteins that may be undergoing positive selection, Jacquard clusters of predicted proteins encoded in seven genomes of *Geobacteraceae* were created using the SYBIL system for comparative genome analyses. Clusters containing at least five predicted proteins were examined through a pipeline that 1) generated protein alignments 2) created nucleotide alignments of the coding DNA from the protein alignments 3) created a phylogenetic tree based on the sequences in the alignment and 4) ran the PAML program to find sites with significant probability of being under positive selection. Only cases in which the minimum protein length in the alignment was >80% of the maximum were used in order to ensure nearly full length homology. Of the 3,201 clusters considered, 38 had genes containing sites with evidence of significant positive selection. Classification of the clusters by biological role category revealed that the largest numbers were annotated as conserved hypothetical proteins (8). The other top role categories in terms of membership were transport and binding proteins (6), enzymes of unknown function (5), energy metabolism (4), protein fate (4) and regulatory function (4). An unusual finding was the presence of 200 sites identified in the essential cell division gene, *fts*Q. Additional analyses of this data is now under way to determine (where protein structures exist) at what locations in the 3D structure of the protein identified sites of positive selection occur and to examine additional data sets for positive selection based on different criteria for cluster membership.

83

GTL

Experimental Mapping and Active Annotation of Transcription Initiation Sites of *Geobacter sulfurreducens*

Katy Juárez[1]* (katy@ibt.unam.mx), Alfredo Mendoza,[1] Leticia Olvera,[1] Maricela Olvera,[1] Enrique Morett,[1] and **Derek R. Lovley**[2] (dlovley@microbio.umass.edu)

[1]Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, México and [2]Department of Microbiology, University of Massachusetts, Amherst, Massachusetts

Project Goals: The main goal of this project is to experimentally map as many transcription start sites (TSS) as possible of genes for which no TSS have yet been described in the literature, using a high throughput technology, and performing a simultaneous active annotation. A database will be generated that will help to better understand promoter structure and gene expression regulation in the *Geobacteraceae*.

Geobacteraceae is the predominant group of bacteria during in situ bioremediation of uranium-contaminated groundwater and on the surface of electrodes harvesting energy from a variety of organic sources. The completion of genome-sequence from multiple members of this group provides the opportunity to apply experimental and computational analyses to obtain new insights into their regulatory processes. Although the availability of the genome sequence of *G. sulfurreducens* has made it possible to predict operon structure, promoters and transcriptional factors binding sites, only a few have been experimentally determined. While such predictions are based mainly on what is known for other bacteria, the

* Presenting author

experimental determination of an important set of transcription start sites will contribute not only to improve the knowledge about the promoter and operon structure in this bacterium, but also increase the predictive capacity for those cases not experimentally determined.

It is well known that the transcription initiation is a critical step in the regulation of gene expression. To know the transcription start sites (TSS) provides relevant information about: a) promoter sequence(s), b) the putative sigma factor(s) and regulatory protein(s) involved in their expression, c) the putative regulatory signals, d) the identity of the initial nucleotide, and also could be used to corroborate operon structure prediction. In order to obtain a global picture of the active promoters in *G. sulfurreducens*, we will follow the large scale TSS mapping methodology that we have previously developed for *E. coli*, which has allowed us to map several hundreds TSS in this organism. To exhaustively map the TSS we will implement two approaches: Directed mapping using a modification of the 5'RACE protocol, and global mapping of TSS using pyrosequencing technology. The modified 5'RACE protocol allowed us to experimentally map more than 300 new TSS in *E. coli*. As controls we mapped the TSS of about 30 genes previously determined experimentally by other groups, confirming the accuracy of our methodology. Interestingly, in this small collection we identified additional TSS for several of these genes not previously detected in the original reports.

We have already tested this methodology in *G. sulfurreducens* and have reported some new TSS, demonstrating that it also efficiently works in this organism. Our strategy for global TSS mapping in *Geobacter* will be to analyze microarray mRNA expression as a guide to select the growth conditions that maximize the expression of as many specific genes and operons as possible. We will start by using the first gene of the predicted operons to map the TSS, we will also analyzed directly the TSS of those genes that have been demonstrated to be physiologically relevant for electron transfer and Fe(III) reduction.

Recently, ultra-high throughput DNA sequencing, using pyrosequencing technology, has allowed rapid determination of the DNA sequence of millions of nucleotides per hour. This technology, implemented in the GS20 instrument of 454 Life Sciences, is based on the sequencing of small individual DNA fragments that are clonally amplified in a very original emulsion format before being sequenced. We have very recently implemented a variation of the standard pyrosequencing methodology to sequence cDNA fragments up to their 3' end. This information will be integrated in a reliable database of promoters and transcription start sites of *G. sulfurreducens* that will ultimately be linked to analyses of transcriptional regulatory systems conforming a database that can be easily accessed.

84

GTL

Evolution of Electron Transfer Out of the Cell: Comparative Genomics of Nine *Geobacteraceae* Genomes

Jessica E. Butler* (jbutler@microbio.umass.edu), Nelson D. Young,* and **Derek R. Lovley** (dlovley@microbio.umass.edu)

Department of Microbiology, University of Massachusetts, Amherst, Massachusetts.

Project Goals: See goals for abstract 79.

Geobacter species grow by transferring electrons out of the cell. In the natural environment, this ability presumably arose to exploit abundant but insoluble electron acceptors like Fe(III) oxide. Fortuitously, this same metabolism can be directed to transfer electrons onto energy-harvesting electrodes, or to reduce and remediate subsurface contaminants like uranium.

Previous research on one species, *Geobacter sulfurreducens*, has shown the importance of periplasmic and outer-membrane *c*-type cytochromes in growth by extracellular electron transfer. Previous research on *G. sulfurreducens* has also shown that complete oxidation of fermentation intermediates like acetate via the TCA cycle is the source of these electrons. The work presented here uses the genome sequences of relatives of *G. sulfurreducens* for three related goals: to assess the conservation of previously studied proteins; to identify the other proteins required to form a complete pathway of electron transfer outside the cell; and to model the evolution of this unique metabolic ability.

Completed or 10x-draft genome sequences are available for *G. sulfurreducens*, *G. metallireducens*, *G. uranireducens*, *G. bemidjiensis*, *G.* strain FRC-32, *G. lovleyi*, *Pelobacter propionicus*, *P. carbinolicus*, and *Desulfuromonas acetoxidans*. For all of these genomes, gene conservation was predicted using two methods: pairs of orthologous proteins were identified as reciprocal best BLAST matches in an all-against-all comparison of two genomes. Their conservation level was quantified in terms of normalized bit scores, and the selective pressure on them in terms of mutation rates. Orthologous proteins from all genomes were grouped using the Markov cluster algorithm. Genes

* Presenting author

acquired by lateral gene transfer from a source outside the family were also identified. A neighbor-joining phylogeny was inferred for every gene in each genome, and those genes that had either a strongly supported atypical phylogeny or a weakly supported atypical phylogeny and out-of-family best BLAST matches were considered foreign.

There are two different types of metabolism in species related to *G. sulfurreducens*. The *Geobacter* and *Desulfuromonas* species are respiratory – they generate ATP by oxidative phosphorylation and the proton motive force. The *Pelobacter* species are fermentative – they oxidize substrates like acetoin to acetate and generate ATP via substrate-level phosphorylation. We were interested in whether the ability to grow by extracellular electron transfer arose from a fermentative or a mixed-metabolism ancestor. A phylogeny of the family was constructed using proteins that were present in a single copy in each genome. These proteins were concatenated, aligned, and phylogeny was modeled using a Bayesian algorithm. The resulting tree showed that there are two clades of *Geobacteraceae*, with a fermentative *Pelobacter* species in each clade. Since phylogeny alone was insufficient to clarify the ancestral metabolism, we examined the conservation of central metabolism and energy metabolism genes in each of the genomes.

In *G. sulfurreducens*, eight reactions of the TCA cycle oxidize acetate and produce NADH, NADPH, and ferredoxin. An NADH dehydrogenase pump protons across the inner membrane for ATP generation via an ATP synthase. Cytochromes transfer the electrons from the reducing equivalents out to extracellular acceptors. All of the enzymes for acetate transport and oxidation are conserved in all of the completed *Geobacter* genomes, as well as *Desulfuromonas*. One of the NADH dehydrogenases and the ATP synthase were among the very best conserved proteins across the six *Geobacter* species. These enzymes had higher averaged conservation scores than many essential housekeeping genes. Also among the best conserved enzymes in all six *Geobacter* species were several TCA cycle enzymes: the 2-oxoglutarate: ferredoxin oxidoreductase, the succinate dehydrogenase, and the citrate synthase.

The fermentative species have lost several key energy metabolism enzymes. Both *Pelobacter* genomes lack an ortholog for the NADPH-generating step of the TCA cycle, the isocitrate dehydrogenase. In addition, both genomes lack the NADPH dehydrogenase. Both *Pelobacter* genomes have also lost a critical inner membrane enzyme, the cytochrome reductase. All of the genomes of the respiratory species contain at least one cytochrome *bc* complex that is predicted to transfer electrons from the quinones in the inner membrane out to the periplasmic cytochromes. Neither *Pelobacter* species contains any genes predicted to fill this role, which would leave them unable to transfer electrons to cytochromes. This indicates why these species are unable to directly reduce Fe(III) or to transfer electrons onto an electrode. The genes required for the fermentation pathways in these species include many that are not found in any of the other family members and are most similar to distantly related species. Thus, it appears that the fermentative species evolved from a respiratory ancestor that gained fermentation genes via lateral transfer and lost the ability to transport electrons via cytochromes outside the cell.

The best conserved proteins across all *Geobacter* species included enzymes for acetate oxidation, proton transport, and ATP synthesis. Surprisingly, none of the cytochromes shown to be required for Fe(III) reduction in *G. sulfurreducens* are highly conserved in all the *Geobacter* species. OmcB, a well-studied outer membrane cytochrome, appears to have orthologs in all species based on clustering and genome context, but the sequence conservation is very low, and the number of hemes bound varies. Most of the other cytochromes shown to be important in Fe(III) reduction in *G. sulfurreducens* are not found in all of the *Geobacter* species, including OmcS, OmcT, OmcZ, and MacA. However, the tri-heme periplasmic cytochrome PpcA appears to be important in all of the *Geobacter* species. Orthologs with moderate levels of sequence conservation are found in several copies in all of the genomes. In addition, there are several cytochromes that are polymers of PpcA-like subunits that are found in all of the *Geobacter* species. Thus, most of the key outer membrane cytochromes are not conserved across all species, while the predominant periplasmic cytochrome is. Broader analysis of cytochromes showed that the best conserved cytochromes include many predicted to be localized in or associated with the inner membrane. These include: the four *c*-type cytochromes encoded near the cytochrome *bc* complexes, and a nine-heme cytochrome with seven transmembrane helices, and one subunits of a very well conserved electron transport complex.

Lack of conservation of key cytochromes indicates that one specific pathway of cytochromes is not required for electron transport out of the cell. However, while specific cytochromes may not be conserved, an abundance of cytochromes is a trait of every *Geobacter* genome. All seven of the respiratory species contain at least 100 cytochromes genes. *G. uranireducens* encodes 152 cytochromes, with a total of 1041 hemes, up to 43 per pro-

* Presenting author

tein. Cytochromes appear to be under less selective pressure than other energy metabolism genes. Comparison of *G. sulfurreducens* to *G. bemidjiensis* showed that 95% of the cytoplasmic and inner membrane energy metabolism genes had orthologs, compared to 46% of multi-heme cytochromes. On average, the selection pressure for the conserved cytochromes was 57% of that for the energy metabolism genes. At the extreme, the outer membrane cytochrome OmcS was under 13% of the pressure on the acetate kinase.

These results indicate that the ancestor of this family was respiratory, not fermentative, oxidized acetate, generated ATP by oxidative phosphorylation at the inner membrane, and contained abundant cytochromes for electron transport out of the cell. These results suggest that further study of the cytochrome(s) that provide the essential electric connection between the inner membrane and periplasm is warranted. However, it appears that once protons are moved across the inner membrane allowing ATP synthesis, the path of electrons across the periplasm and outer membrane is likely very variable, and possibly non-specific.

85

GTL

Molecular Analysis of the Metabolic State of *Geobacter* Species During in Situ Uranium Bioremediation

D.E. Holmes,* P.J. Mouser, H. Elifantz, L.A.L. N'Guessan, M. Aklujkar, M.A. Chavan, and **D.R. Lovley** (dlovley@microbio.umass.edu)

Department of Microbiology, University of Massachusetts, Amherst, Massachusetts

Project Goals: See goals for abstract 79.

Analysis of 16S rRNA sequences, lipids, metagenomic sequences, and environmental proteomics have all unequivocally demonstrated that *Geobacter* species are the predominant organisms catalyzing *in situ* uranium bioremediation at the DOE study site at Rifle, CO. Numerous other studies using one or more of these methods have also suggested that *Geobacter* species are the most abundant dissimilatory metal-reducing microorganisms in a diversity of other contaminated subsurface environments. Therefore, a primary goal of our Genomics:GTL project is to develop genome-based strategies that will make it possible to: 1) elucidate the metabolic state of *Geobacter* species in the subsurface in order to determine factors that might be limiting the bioremediation process and 2) develop genome-scale *in silico* models that can predict how the metabolism and growth of *Geobacter* species will respond to changes in environmental conditions resulting from geochemical heterogeneities and/or environmental changes that are imposed with the goal of optimizing *in situ* bioremediation.

In the last year significant progress has been made in elucidating mechanisms behind the physiological responses of the natural community of *Geobacter* species in the subsurface to important environmental changes. For example, nitrogen is a key nutrient for *Geobacter* species in the subsurface. Our previous analysis of gene transcript abundance in the subsurface during *in situ* uranium bioremediation at the Rifle, CO site demonstrated that *in situ Geobacter* species were limited for ammonium and thus had to fix nitrogen from the atmosphere. In order to investigate this phenomenon further, expression of genes for nitrogen fixation and ammonium uptake was evaluated at a series of geochemically diverse wells representing ammonium concentrations ranging from 0-400 μM during the 2007 field experiment at the Rifle site. In one monitoring well in which ammonium concentrations were below detection, the number of *Geobacter* transcripts from *amtB*, which encodes an ammonium transport protein, increased 25 fold when acetate was added to stimulate the growth of *Geobacter* and U(VI) reduction. Transcript levels for *nifD*, a gene involved in nitrogen fixation also increased 27 fold.

Pure culture studies with *Geobacter sulfurreducens* identified *pstB*, which encodes a phosphate transporter, and *phoU*, which encodes a phosphate uptake regulatory protein, as key genes with increased expression during phosphate limitation. Preliminary analysis of gene transcript levels in the subsurface during *in situ* uranium bioremediation suggested that the natural community of *Geobacter* had a significant phosphate requirement as growth was initially stimulated, but that later in the bioremediation process *Geobacter* were probably limited by the availability of other nutrients.

Studies with *G. sulfurreducens* and *G. uraniumreducens*, which was isolated from the Rifle site, identified genes encoding putative acetate transporters that were expressed at high levels when acetate availability limited growth. Transcript levels for these transporters in the natural community of *Geobacter* species were monitored during the course of a field experiment in which the acetate additions were manipulated to alter acetate availability in the groundwater. In accordance with pure culture studies, transcript levels for the acetate transporter genes decreased ca. 20-fold as high (≥ 3 mM) concentrations of acetate became available in the groundwater.

* Presenting author

When acetate additions were temporarily decreased, reducing groundwater concentrations below 1 mM, transcript levels again increased ca. 8-fold.

In order to determine the response of *in situ Geobacter* to various nutrient amendments, molecular tools that can be used to monitor changes in the growth rate of *Geobacter* in the environment were developed. Microarray analyses of *G. sulfurreducens* and *G. uraniumreducens* cells grown at different growth rates identified several genes whose expression was directly correlated with cell growth rates. These genes included the ribosomal proteins, *rplL* and *rpsC*, and the cell division proteins, *ftsA* and *ftsZ*. Expression of these genes by *G. uranireducens* was then monitored over time during growth in sediments and on the soluble electron acceptor, fumarate, at 18°C, 30°C, and 37°C. Growth of *G. uranireducens* in sediments was significantly slower at higher temperatures, and the number of mRNA transcripts from all 4 of these genes at mid log phase was also lower at the higher temperatures. In addition, the number of mRNA transcripts from *ftsZ* and *rpsC* corresponded with the number of cells and acetate concentrations in the fumarate-grown cultures. Similar results were observed in groundwater collected during the uranium bioremediation field experiment.

In order to refine the existing *Geobacter* genome-based *in silico* model, the genes that are found in the *Geobacter* species that predominate in subsurface environments were further investigated. This was accomplished via: 1) isolation of several new *Geobacter* strains representative of the *Geobacter* species that predominate in the subsurface; 2) direct sequencing of genomic DNA collected during *in situ* uranium bioremediation; and 3) amplifying and sequencing genomic DNA from single cells isolated from the subsurface. Phylogenetic analysis of this genome data indicated that *Geobacter* species at these sites typically cluster into a tight phylogenetic group, referred to as Subsurface Clade 1. Several subsurface clade 1 genomes have been completed or are in the process of being sequenced. Analysis of these genomes has indicated that they all have complete TCA cycles and numerous c-type cytochrome genes. Analysis of the complete *G. uraniumreducens* genome showed that it is significantly larger than those of other *Geobacteraceae*, and is marked by an abundance of mobile genetic elements (~3% of genes) and other features indicative of active gene exchange, duplication, rearrangement and loss. Large insertion/deletion events in flagellar structural genes could explain the long flagella of *G. uranireducens*.

The genomes of *G. bemidjiensis* and strain M21 are particularly similar, even though they were recovered from two geographically and geochemically distinct subsurface environments. *G. bemidjiensis* was isolated from a petroleum contaminated aquifer located in Bemidji, MN, while strain M21 was isolated from the uranium-contaminated aquifer in Rifle, CO. Comparison of *G. bemidjiensis* and strain M21 genomes revealed that more than half of the c-type cytochromes present in *G. bemidjiensis* have homologues in strain M21. In addition, unlike other *Geobacter* species, the genomes of *G. bemidjiensis* and strain M21 contain *E. coli*-like glucokinase, galactokinase, galactose-1-phosphate uridylyltransferase, and hydroxypyruvate reductase genes which are all involved in glucose and galactose catabolism. In addition, the glucose/galactose transporter was detected in *G. bemidjiensis*. Preliminary analysis of metagenomic and single cell data showed that the majority of sequences in these libraries are most similar to *G. bemijiensis* and strain M21.

86

GTL

Molecular Mechanisms Regulating Gene Expression in *Geobacter sulfurreducens* Under Environmentally Relevant Conditions

Toshiyuki Ueki[1]* (tueki@microbio.umass.edu), Ching Leang,[1] Maddalena V. Coppi,[1] Regina A. O'Neil,[1] Kelly P. Nevin,[1] Trevor L. Woodard,[1] Anna Liu,[2] Barbara A. Methé,[3] and **Derek R. Lovley**[1] (dlovley@microbio.umass.edu)

[1]Department of Microbiology and [2]Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts and [3]J. Craig Venter Institute, Rockville, Maryland

Project Goals: See goals for abstract 79.

In order to predictively model the activities of *Geobacter* species during *in situ* uranium bioremediation and on the surface of electrodes harvesting electricity from waste organic matter and renewable biomass, we have been investigating the molecular mechanisms regulating gene expression under various environmental conditions in a representative of *Geobacter* species, *Geobacter sulfurreducens*. This information is essential for improved *in silico* models that are being employed in the optimization of current production and bioremediation practices.

The number of homologs of regulatory proteins such as transcription factors and signal transduction proteins is large in order for organisms to promptly and properly sense and respond to a variety of environmental changes.

* Presenting author

Therefore, it is important to understand the function of each homolog. It is, however, necessary to develop a system-level analysis for the elucidation of regulatory networks in global gene expression. Such an analysis can be integrated by functional (e.g. microarray analysis) and comparative genomics. We are developing a systematic analysis to elucidate gene regulation in *Geobacter* species, as microarray analyses under a variety of conditions and genome sequences of many *Geobacter* species are available. For example, we recently applied a systematic analysis integrated by functional and comparative genomics in combination with biochemical and genetic methods to elucidate gene regulation during nitrogen fixation. This is of interest because our previous studies have demonstrated that *Geobacter* species are forced to utilize atmospheric nitrogen as a nitrogen source during in situ uranium bioremediation. We identified novel regulatory cascades controlling gene expression during nitrogen fixation in *G. sulfurreducens.* The cascades consisted of the two two-component systems GnfL/GnfM and GnfK/GnfR. The GnfL/GnfM system appeared to be the master regulator for the expression of genes involved in nitrogen fixation. It is likely that the GnfK/GnfR system was under the control of the GnfL/GnfM system and regulated a subset of genes by antitermination. Mutants of GnfK and GnfR were able to grow in the presence of fixed nitrogen, while they were unable to grow in the absence of fixed nitrogen. Such a systematic analysis integrated by functional and comparative genomics in combination of biochemical and genetic methods should also be applicable to the elucidation of other gene regulation in *G. sulfurreducens* as well as other *Geobacter* species.

The two-component system is an important strategy for sensing and responding to environmental conditions. It typically consists of a sensor kinase, which senses an environmental signal, and a response regulator, which regulates gene expression to adapt to an environmental change. The genome of *G. sulfurreducens* encodes an unusually large number of proteins belonging to the two-component system, which may reflect the need to adapt to a myriad of different conditions in subsurface environments. The two-component system GsuTCS1 was found to be involved in Fe(III) reduction as well as redox sensing. The sensor kinase of GsuTCS1 has a unique sensor domain containing *c*-type heme binding motifs. To further understand the function of GsuTCS1, a strain over-expressing the sensor kinase and the response regulator of GsuTCS1 was constructed. This strain showed biofilm formation under a condition where the wild-type strain does not form biofilms, suggesting that GsuTCS1 regulates genes involved in biofilm formation. Biofilm formation has been shown to be an important cellular process for optimal current production in microbial fuel cells.

Our previous studies have demonstrated that *in situ* transcript levels of the gene for citrate synthase in groundwater during *in situ* uranium bioremediation can be used to infer rates of *Geobacter* metabolism in the subsurface. Citrate synthase is a key enzyme in the TCA cycle and regulates the entry of carbon into the TCA cycle. Investigations into the mechanisms regulating expression of *gltA*, the citrate synthase gene, demonstrated that the primary transcriptional regulator was a repressor. Genome-wide sequence analysis identified additional genes, whose putative promoters contain a DNA element with high similarity to the repressor binding site in the *gltA* promoter and which are likely involved in biosynthesis and energy generation, in *G. sulfurreducens* as well as other *Geobacter* species. This suggests that the transcriptional repressor involved in *gltA* expression is a global regulator for biosynthesis and energy generation among *Geobacter* species. By a systematic analysis combined with comparative genomics and a biochemical assay, several transcription factors were identified as candidates for the repressor. These insights into the regulation of central metabolisms provide important information for modeling the cellular activities of *Geobacter* species.

G. sulfurreducens has previously been considered a strict anaerobe. However, it has been shown that *G. sulfurreducens* is able to tolerate exposure to atmospheric oxygen and to grow with oxygen as the sole electron acceptor. These results may explain how *Geobacter* species rapidly become the predominant Fe(III)-reducing microorganisms in a diversity of subsurface sediments, when organic electron donors are introduced into previously oxic subsurface sediments in order to stimulate anaerobic dissimilatory metal reduction. Two genes encoding putative terminal oxidases, cytochrome *c* oxidase and cytochrome *bd* ubiquinone oxidase, were identified in the *G. sulfurreducens* genome. Cytochrome *bd* ubiquinone oxidase was found to play a key role in oxygen metabolism. The expression of the gene encoding cytochrome *bd* ubiquinone oxidase was under the control of multiple factors such as the stationary-phase sigma factor RpoS, FNR homologs and PerR homologs. These findings add significant knowledge to our understanding of the molecular strategies, which *Geobacter* species utilize to survive oxic environments that they encounter in the subsurface.

One of the defining features of *Geobacter* species is a three gene cluster consisting of two genes encoding Fe(II)-dependent transcriptional repressors, *fur* and *ideR*, flanking a gene for a ferrous uptake protein, *feoB-1*. Because *Geobacter* species can be exposed to a wide range

* Presenting author

of Fe(II) concentrations in the subsurface and have unusually high assimilatory iron requirements due to their exceptionally high *c*-type cytochrome content, tight control over intracellular iron homeostasis is likely to be critical for their survival. Phenotypic and microarray analyses with *G. sulfurreducens* mutants lacking *fur* and *ideR* indicated that both Fur and IdeR function as Fe(II)-dependent transcriptional repressors that simultaneously regulate genes involved in intracellular iron homeostasis, protein folding and metabolisms. Thus, these genes may play a critical role in allowing *Geobacter* species to respond to and thrive in fluctuating Fe(II) concentrations in the subsurface.

Our understanding of regulatory mechanisms at the molecular levels in *Geobacter* species has been expanding in quantity as well as quality. These studies allow us to improve our models for predicting how *Geobacter* species respond to various changes in environmental conditions during *in situ* bioremediation and energy harvesting.

87

GTL

Towards a Transcription Regulatory Network (TRN) in *Geobacter sulfurreducens*

Karsten Zengler[1]* (kzengler@bioeng.ucsd.edu), Yu Qiu,[1] Byung-Kwan Cho,[1] Larisa Kagan,[1] James Elkins,[1] **Derek R. Lovley**,[2] and Bernhard Palsson[1]

[1]University of California, San Diego, California and [2]University of Massachusetts, Amherst, Massachusetts

A comprehensive study of genome-wide protein-DNA interactions was performed with *E. coli* as a model organism. Chromatin immunoprecipitation (ChIP) coupled with high-density tiling arrays (ChIP-chip) has been applied to perform genome-wide location analysis for three classes of DNA proteins in *E. coli*. Proteins examined were RNA polymerase and sigma factors, broad acting transcription factors as well as DNA bending proteins. The genome-scale transcriptional regulatory network established for *E. coli* allows for the first time in depth understanding of fundamental aspects of chromosome structure, DNA replication, DNA repair, response to stress and regulation of metabolism.

Towards a comprehensive understanding of the transcriptional regulatory network in *Geobacter sulfurreducens* we adapted the protocol established in our laboratory for *E. coli* to work in *G. sulfurreducens*. *G. sulfurreducens* is capable of transferring electrons to a variety of electron acceptors including Fe(III), U(IV), and the surface of electrodes making it the candidate of choice for bioremediation of contaminated environments and harvesting electricity from waste organic matter. In depth understanding of how *G. sulfurreducens* functions will have a great impact on optimizing bioremediation and energy harvesting applications. Genome-wide binding patterns of RNA polymerase and σ^{70} in *G. sulfurreducens* were determined by ChIP-chip analysis. Patterns of RNA polymerase and σ^{70} binding were compared from cells grown with acetate under fumarate- and Fe(III)-reducing conditions. All ChIP-chip results were complemented by gene expression profiles using high-density tiling arrays. In addition promoter regions in *G. sulfurreducens* were determined under various growth conditions by ChIP-chip analysis of cells treated with rifampicin (rifampicin inhibits transcription elongation by RNA polymerase but not its binding to promoter region). Polyclonal antibodies against further sigma factors from *G. sulfurreducens* (RpoE, RpoH, RpoN, and RpoS) are currently produced. In addition, we established a system for routine epitope-tagging of transcription factors in *G. sulfurreducens*.

88

GTL

Structural Characterization of *Geobacter sulfurreducens* Heme Proteins: Two Novel Periplasmic Sensor Domains from Chemotaxis Proteins, and the Soluble Part of a Membrane Protein OmcF

P. Raj Pokkuluri,[1] Yuri Y. Londer,[1] Norma Duke,[1] Stephan Wood,[1] Miguel Pessanha,[2] Leonor Morgado,[2] Teresa Catarino,[3] Carlos A. Salgueiro,[2] and **Marianne Schiffer**[1]* (mschiffer@anl.gov)

[1]Biosciences Division, Argonne National Laboratory, Argonne, Illinois; [2]Requimte, Centro de Química Fina e Biotecnologia, Dep. Quimica, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal; and [3]Instituto de Tecnologia Quimica e Biologica, Universidade Nova de Lisboa, Oeiras, Portugal

Project Goals: As sub-project of GTL grant "Genome-based models to optimize in situ bioremediation of uranium and harvesting electrical energy from waste organic matter, Derek Lovley (PI)" our goals are to analyze selected proteins to understand their function in the cell. This includes modeling of structures based on their amino acid sequences, determination of their

* Presenting author

structures, and the functional interpretation of the structures, such as active sites and surface properties.

Geobacter sulfurreducens encodes over 100 cytochromes containing *c*-type hemes. We determined the structures of periplasmic sensor domains of two methyl-accepting chemotaxis proteins encoded by GSU0582 and GSU0935 (1) and the structure of the soluble part of a membrane protein OmcF encoded by GSU2432 (2).

The heme containing sensor domains (about 135 residues) were expressed in *E. coli*, characterized in solution, and their crystal structures were determined. The R factor is 24.7% for 2.0 Å data and 20.5% for 1.9 Å for GSU0582 and GSU0935 sensors, respectively. In the crystal, both sensor domains form "swapped" dimers and reveal a novel way of forming PAS-type domains using two chains. These are the first PAS domain structures that contain a covalently bound heme. The swapped segment consists of two helices of about 45 residues at the N-terminus with the hemes located between the two monomers. In the case of GSU0582 sensor, the dimer is related by a crystallographic 2-fold axis and the heme is coordinated by an axial His and a water molecule. In the case of GSU0935 sensor, the crystals contain a non-crystallographic dimer, and surprisingly, the coordination of the heme in each monomer is different; Monomer A heme has His-Met ligation and Monomer B heme has His-water ligation as found in the GSU0582 sensor. Optical absorption, EPR and NMR spectroscopies have revealed that the heme groups of both sensor domains are high- and low-spin in the oxidized and reduced forms, respectively, and that the spin-state interconversion involves a heme axial ligand replacement. The reduction potentials of the sensor domains of GSU0582 and GSU0935 are -156 mV and -251 mV, respectively.

At present, we do not know what compound or compounds these chemotaxis proteins sense in *G. sulfurreducens* but homologous proteins are conserved in the *Geobacteraceae* family. Even though the two sensors have similar structures and spectroscopic properties, the reduction potentials of the heme groups are quite distinct, differing by approximately 100mV, suggesting that they might perform their physiological functions in environments with different redox potentials. We propose that these periplasmic sensor domains could be part of a global cellular mechanism for sensing the periplasmic redox potential or small ligands. The swapped dimerization of these sensor domains and redox-linked ligand switch might be related to the mechanism of signal transduction by these chemotaxis proteins. We suggest that the helix swapped dimer formation could be a mechanism for signal transduction from the periplasm to the cytoplasm as the swapped dimer will alter the relative positions of the trans-membrane helices that can lead to a change in the effector part of the molecule located in the cytoplasm.

OmcF encoded by GSU2432, was identified as an outer-membrane heme protein that affects the expression of the membrane proteins, OmcB and OmcC, in *G. sulfurreducens* (2). The OmcF deficient strain is impaired in its ability to grow on Fe(III) citrate (2). OmcF is a 104 residue protein with one heme attachment site and a lipid anchor. The soluble part of OmcF is homologous to cytochromes c_6 found in algae and cyanobacteria, where the cytochrome c_6 transfers electrons from cytochrome b_6f to photo system I. The soluble part of OmcF, from residue 20, was expressed in *E. coli* and its structure was determined by X-ray diffraction; the R-factor is 17.0% for 1.35 Å data. It is a protein with mostly helical structure. The heme is His - Met coordinated. The c_6 from green algae *Monoraphidium braunii* (3) is a very close structural homolog with a DALI score, Z=13.1. The reduction potential of OmcF is about +129 mV; interestingly the midpoint potential of *M. braunii* c_6 is +358 mV (4).

X-ray diffraction data were collected at Structural Biology Center beam line of the APS.

References

1. Londer YY, Dementieva IS, D'Ausilio CA, Pokkuluri PR and Schiffer M (2006). Characterization of a *c* type heme containing PAS sensor domain from *Geobacter sulfurreducens* representing a novel family of periplasmic sensors in *Geobacteraceae* and other bacteria. *FEMS Microbiol Lett* 258: 173-181.
2. Kim B-C, Leang C, Ding Y-HR, Glaven RH, Coppi MV and Lovley DR (2005). OmcF, a Putative *c*-Type Monoheme Outer Membrane Cytochrome Required for the Expression of Other Outer Membrane Cytochromes in *Geobacter sulfurreducens*. *J. Bacter.* 187: 4505-4513.
3. Frazo C, Soares CM, Carrondo MA, Pohl E, Dauter Z, Wilson KS, Hervais M, Navarro JA, De la Rosa MA and Sheldrick GM (1995). Ab initio determination of the crystal structure of cytochrome c_6 and comparison with plastocyanin,. *Structure*, 3: 1159-1169.
4. Campos AP, Aguiar AP, Hervas M, Regalla M, Navarro JA, Ortega JM, Xavier AV, De La Rosa MA and Teixeira M (1993). Cytochrome c_6 from *Monoraphidium braunii*. A cytochrome with an unusual heme axial coordination. *Eur. J. Biochem.* 216: 329-341.

* Presenting author

89

GTL

Dynamic Systems-Level Analysis of Oxygen-Dependent Cell State Transitions Reveals a Surprising Chronology of Cellular Events

Amy K. Schmid,[1] David J. Reiss,[1] Amardeep Kaur,[1] Min Pan,[1] Nichole King,[1] Phu T. Van,[1] Laura Hohmann,[1] Daniel B. Martin,[1,2] and **Nitin S. Baliga**[1]* (nbaliga@systemsbiology.org)

[1]Institute for Systems Biology, Seattle, Washington and [2]Divisions of Human Biology and Clinical Research, Fred Hutchinson Cancer Research Center, Seattle, Washington

Adjustment of physiology in response to changes in oxygen availability is critical for the survival of all organisms. However, the chronology of events and the regulatory processes that determine how and when changes in environmental oxygen tension result in an appropriate cellular response is not well understood at a systems level. Therefore, transcriptome, proteome, ATP, and growth changes were analyzed in a halophilic archaeon to generate a temporal model that describes the cellular events that drive the transition between the organism's two opposing cell states of anoxic quiescence and aerobic growth. According to this model, upon oxygen influx, an initial burst of protein synthesis precedes ATP and transcription induction, rapidly driving the cell out of anoxic quiescence, culminating in the resumption of growth. This model also suggests that quiescent cells appear to remain actively poised for energy production from a variety of different sources. Dynamic temporal analysis of relationships between transcription and translation of key genes suggests several important mechanisms for cellular sustenance under anoxia as well as specific instances of posttranscriptional regulation. Importantly, genes which code for protein complexes involved in energy generation appear to be regulated in tandem. For instance, aerobic metabolism complexes such as the ribosome, ATP synthase, and succinate dehydrogenase appear to have similar time lags between transcription and translation. Time lagging analysis of anaerobic energy generation complexes is currently underway.

90

GTL

Rhythmic Gene Expression in *Halobacterium* NRC-1 After Day/Night Cycle Entrainment

Kenia Whitehead,[1] Min Pan,[1] Kenichi Masumura,[2] Tetsuya Mori,[3] Lena Suzuki,[3] Carl. H. Johnson,[3] Richard Bonneau,[4] and **Nitin S. Baliga**[1]* (nbaliga@systemsbiology.org)

[1]Institute for Systems Biology, Seattle, Washington; [2]Division of Genetics and Mutagenesis, National Institute of Health Sciences, Tokyo, Japan; [3]Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee; and [4]Center for Genomics and Systems Biology, New York University, New York, New York

Small protein complexes forming biological oscillators are used by many organisms to coordinate gene expression, metabolism and cell physiology. Thus far, molecular oscillators for day/night cycles have not been identified in prokaryotes with the sole exception of cyanobacteria. Our goal is to study day/night rhythms in gene expression in the halophilic archaeon, *Halobacterium* NRC-1 during and after entrainment. Recent results have shown that ~10% of the genome follows a pattern mediated by light/dark cycles with peaks in gene expression at the day/night transition even after the day/night stimulus has been removed and cells are in constant darkness. Rhythmically expressed genes encode enzymes at key positions in carbon and nitrogen metabolism as well as a host of transcriptional regulators. While the molecular basis for this rhythm generation and memory in *Halobacterium* NRC-1 is unknown we can now begin to examine the protein complexes mediating rhythm generation in archaea.

* Presenting author

91

GTL

Transcriptome Structure of *Halobacterium sp.* NRC-1 and Its Dynamic Changes During Growth

Tie Koide,[1] Marc Facciotti,[1] David Reiss,[1] Kenichi Masumura,[2] Christopher Bare,[1] Min Pan,[1] and **Nitin Baliga**[1]* (nbaliga@systemsbiology.org)

[1]Institute for Systems Biology, Seattle, Washington and [2]Division of Genetics and Mutagenesis, National Institute of Health Sciences, Tokyo, Japan

Project Goals: MAGGIE integrates teams at Lawrence Berkeley National Lab and the Advanced Light Source (ALS) with researchers at the Scripps Research Institute, the University of Georgia, the University of California Berkeley, and the Institute for Systems Biology to achieve a molecular-level understanding of the dynamic macromolecular machines that underlie all of microbial cell biology. MAGGIE is providing improved 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 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.

Microarray analysis of transcriptome changes in *Halobacterium salinarum sp. NRC-1* have revealed that nearly half of all annotated genes undergo some significant change in expression to help mediate the physiological transitions that occur during grown. In most cases, dynamic assemblies of molecular protein complexes at key genomic loci regulate these changes. In order to understand how these transcriptional complexes contribute to the regulation of gene expression we must first understand precisely when and where they assemble in the genome. One potentially key indicator for localizing complex assembly is to map the transcript boundaries throughout the genome. Using high-resolution tiling arrays, we have constructed a transcriptome structure map generated for *Halobacterium. sp NRC-1* from cells harvested at various optical densities throughout a standard laboratory growth experiment. This has enabled us to accurately map transcript boundaries including features such as alternative transcription start sites and operon boundaries. When integrated with existing ChIP-chip (protein-DNA interaction) data the detailed and dynamic transcriptome map allows for great insight into the assembly of transcriptional complexes and their influence on patterns of global gene expression. Furthermore, we have also revealed the expression of nearly 150 putative non-coding regions, nearly half of which are differentially transcribed during growth. These non-coding RNAs are poorly understood in archaeal systems but have been shown to act as important post-transcriptional regulators in bacteria and eukaryotes by altering the stability of mRNAs, translation initiation and the assembly of protein complexes. The interplay between these putative regulatory RNAs and their targets is currently under investigation.

92

GTL

The *Caulobacter* Divisome: Parts List, Assembly, and Mechanism of Action

Erin D. Goley,[1] Yi-Chun Yeh,[1] Luis Comolli,[2] J.N. Werner,[1] Ken Downing,[2] Zemer Gitai,[3] **Harley H. McAdams**[1] (hmcadams@stanford.edu), and Lucy Shapiro[1]*

[1]Stanford University School of Medicine, Stanford, California; [2]Lawrence Berkeley National Laboratory, Berkeley, California; and [3]Princeton University, Princeton, New Jersey

Project Goals: The objective of the project is development of a global and integrated view of the dynamically changing structures involved in the bacterial cell's function, cell cycle progression, response to environmental challenges, and polar development. We seek to identify multiprotein complexes associated with regulation of bacterial development, chromosome replication and segregation, cytokinesis, and intracellular uranium sequestration, as well as the regulatory processes and protein complexes that construct, position, and degrade these structures.

The cell division apparatus (divisome) in bacteria mediates the constriction of the cell membranes and the inward growth of the cell wall in coordination with cell growth and chromosome segregation. Despite decades of study, however, relatively little is known about the structure and assembly of the divisome or the molecular functions of its components. The tubulin relative FtsZ is the best characterized and most highly conserved divisome protein. It is a GTPase that polymerizes near midcell, defining the site of cell division. FtsZ serves as a scaffold for assembly of the divisome and is hypothesized to generate constrictive forces. As the structure, dynamics and function of FtsZ are likely to be regulated by inter-

* Presenting author

acting partners, we sought to identify all FtsZ-binding proteins in *Caulobacter*. To do this, we developed an assay in which we overproduce a GTPase-defective mutant of FtsZ, causing cells to adopt a distinct morphology wherein long, slender constrictions containing FtsZ separate the cell bodies. We found that fluorescent fusions to *Caulobacter* homologs of all known FtsZ-binding proteins co-localize with FtsZ in the constrictions of mutant cells, whereas divisome proteins that do not bind to FtsZ are diffuse. To identify novel FtsZ-binding proteins we used this assay to probe the library of *Caulobacter* strains bearing fluorescent fusions to 442 different localized proteins generated by the Gitai lab at Princeton. From this screen, we identified six proteins that clearly localized to the constrictions, four of which are previously uncharacterized. To date, we have confirmed that three of these proteins bind directly to FtsZ *in vitro*. Surprisingly, we discovered an additional set of proteins in our screen that were specifically excluded from the FtsZ-rich constrictions, indicating a second mode of FtsZ-directed protein localization in *Caulobacter*. We are now taking genetic, cytological, and biochemical approaches aimed at uncovering the mechanisms by which FtsZ controls the localization of these factors and determining their cellular functions.

In addition to those new FtsZ-binding proteins identified in our screen, at least sixteen other proteins are recruited to form the divisome downstream of FtsZ. To gain molecular insight into their functions, we have cloned the *Caulobacter* homologs of the known cell division proteins and have so far confirmed that thirteen of these are essential for cytokinesis in this organism. Moreover, we find that fluorescent fusions to each of these proteins localize to the division site and we are in the midst of a careful study following the timing of their localization to better understand the process of divisome assembly. Finally, we have made significant progress in understanding the role of the Tol-Pal complex in *Caulobacter* cytokinesis. Using a combination of genetic, cell biological, and high resolution microscopy techniques (in collaboration with the Downing lab at LBNL) we have shown that Pal is specifically required for outer membrane integrity and for invagination of the outer membrane during cytokinesis. Current and future efforts are directed at similarly detailed structural and functional analyses of the other *Caulobacter* divisome proteins.

This work was funded by the DOE GTL project titled "Dynamic spatial organization of multi-protein complexes controlling microbial polar organization, chromosome replication, and cytokinesis." The PI of this multiinstitutional grant is Harley McAdams, Stanford University.

93

GTL

Development of Methods for Correlated Light and Electron Microscopic Analysis of Protein Complexes: Application to *Caulobacter crescentus*

Guido M. Gaietta,[1] Sean Prendergast,[1] Ying Jones,[1] James Obayashi,[1] Thomas J. Deerinck,[1] Grant Bowman,[2] Yi Chun Yeh,[2] Lucy Shapiro,[2] **Harley McAdams**[2] (hmcadams@stanford.edu), and Mark H. Ellisman[1]*

[1]National Center for Microscopy and Imaging Research, Department of Neurosciences, University of California, San Diego, California; [2]Department of Developmental Biology, Stanford University, School of Medicine, Beckman Center, Stanford, California; and [3]Life Science Division, Lawrence Berkeley National Laboratory, University of California, Berkeley, California

Project Goals: The specific aims in the original proposal were: (1) Adapt 4Cys tags and related compounds to the labeling and analysis of protein complexes in *Caulobacter crescentus*. (2) Enhance the throughput of the end-to-end electron tomographic data collection and volume generation process.

Microbial cells present interesting challenges to multiscale microscopic analysis by correlated light and electron microscopic imaging (LM and EM). Challenges include limited permeability of the cell wall, sensitivity of the cells to conventional chemical fixatives and their small size. We have systematically addressed these obstacles and developed methods that provide for high quality ultrastructure and enable correlation of fluorescent signals with electron density of corresponding subcellular structures in 3D electron tomographic reconstructions. These methods are now being extended to allow detection of signals from specific complexes of interest in *Caulobacter crescentus*, as originally delineated in the GTL project *Dynamic spatial organization of multi-protein complexes controlling microbial polar organization, chromosome replication, and cytokinesis*. Specifically:

1. To optimize preservation of supramolecular structure in fluorescently photoconverted specimens we have developed methods that combine fluorescence photoconversion with high-pressure freezing and freeze-substitution (HPF/FS). This allows use of rapid primary aldehyde fixation and labeling of specific protein complexes by fluorescence photocon-

* Presenting author

version with the advantages of HPF, which avoid the bulk specimen shrinkage due to solvent dehydration. Samples prepared for electron tomography and 3D reconstruction using this new approach exhibit spectacular ultrastructure and overall form.

2. To further enhance the detection of lower abundance proteins we are exploring approaches that increase the deposition of electron dense products at sites bearing fluorescent reporters. New methods include the incorporation of multiple tetracysteine motifs to the target protein/fluorescent protein complex and enhancing rates of photoconversion by utilizing more reactive derivatives of the diaminobenzidine chromogen as well as the addition of catalytic cofactors.
3. Recent developments in super-resolution LM have stimulated us to devise alternative approaches that overcome the aforementioned obstacles. These new approaches have been designed to be compatible with methods like photoactivation localization microscopy (PALM), stochastic optical reconstruction microscopy (STORM) or stimulated emission depletion (STED) microscopy, and enable the acquisition of images containing information well beyond the diffraction limit of conventional light optical systems. The key enabling technology is the development of procedures that maintain fluorescent signals in fixed and resin embedded specimens prepared by HPF and epoxy fixation/FS. The ability to simultaneously preserve fluorescent signals and cellular ultrastructure after embedding in epoxy resins represents a powerful alternative to fluorescent photoconversion and should greatly enhance the ability to detect lower abundance proteins. The applicability of this approach is further increased by using fusions with photoactivatable/photoswitchable fluorescent proteins, which retain their function after embedding and allow combined high super-resolution LM protein localization followed by directly correlated electron microscopy and 3D reconstruction using electron tomography.

This work was funded by the DOE GTL project titled "Dynamic spatial organization of multi-protein complexes controlling microbial polar organization, chromosome replication, and cytokinesis". The PI of this multiinstitutional grant is Harley McAdams, Stanford University.

94

GTL

Mapping Protein-Protein Interaction Networks of *Caulobacter crescentus* using Tandem Affinity Purification

Jian Zhu,[1] Ping Hu,[1] Tom Taverner,[2] Gary Andersen,[1] Lucy Shapiro,[2] **Harley McAdams**[2] (hmacadams@stanford.edu), and Thomas Earnest[1]* (TNEarnest@lbl.gov)

[1]Lawrence Berkeley National Laboratory, Berkeley, California and [2]Stanford University, Palo Alto, California

Project Goals: Dynamic localization of multiprotein complexes and their structural studies at molecular and cellular levels.

Tandem Affinity Purification (TAP) has proven to be a powerful tool for the investigation of protein-protein interactions in yeast and in *E. coli* as it can purify stable multi-protein complexes in near native conditions. We have developed methods of integrating the TAP tag with the target ORF at the C-terminal using the Gateway system to allow for higher target throughput to be achieved. Tagging of the N-terminal can be accomplished by a more complicated protocol for targets of high scientific value where C-terminal tagging is not possible, when the tag would interfere with the C-terminal interactions with other proteins required for proper function, or when the C-terminal is cleaved during proteolytic regulation. In our initial test set, which consisted of targets covering a wide range of sizes and functionality, approximately 75% of the targets could be C-terminal TAP tagged and purified through the second affinity column. Investigations of ORFs targeted due to their importance in cell polarity, cell division, and cell cycle control also exhibited a high success rate for tag incorporation and purification. For example the hybrid histidine kinase, CckA, co-purifies with a number of proteins not pulled down from wild-type cells after identical growth, disruption, and purification. Mass spectroscopic analysis of the other proteins that co-purify is underway. Components of multiprotein complexes that are identified by this approach are then over expressed by a variety of methods to produce multi-protein complexes and sub complexes in sufficient quantities for structural studies to be performed.

Funding from the DOE Genomics:GTL project titled "Dynamic spatial organization of multi-protein complexes controlling microbial polar organization, chromosome replication, and cytokinesis." The Principal Investigator of this multi-institutional grant is Harley McAdams, Stanford University.

* Presenting author

95

GTL

A Polymeric Protein Anchors the Chromosome Origin and Governs Multiprotein Complex Assembly at a Bacterial Cell Pole

Grant Bowman,[1] Luis Comolli,[2] Marcelle Koenig,[1] Jian Zhu,[2] Thomas Earnest,[2] Ken Downing,[2] W. E. Moerner,[1] and **Lucy Shapiro**[1]* (shapiro@stanford.edu)

[1]Stanford University, Sanford, California and [2]Lawrence Berkeley National Laboratory, Berkeley, California

Project Goals: The objective of the project is development of a global and integrated view of the dynamically changing structures involved in the bacterial cell's function, cell cycle progression, response to environmental challenges, and polar development. We seek to identify multiprotein complexes associated with regulation of bacterial development, chromosome replication and segregation, cytokinesis, and intracellular uranium sequestration, as well as the regulatory processes and protein complexes that construct, position, and degrade these structures.

In rod shaped bacteria, the cell poles are often utilized as positional reference points for the localization of cellular structures, such as flagella and stalks, polar regulatory complexes of phospho-signaling proteins, and protein complexes that regulate cell division. In some species, the chromosomal origins are also targeted to the cell poles. We find that multiple aspects of cellular organization are linked in *Caulobacter crescentus*, in that they have a common requirement for a novel protein, GmpA. In the absence of GmpA, the chromosomal origins become dissociated from the pole, exhibiting a wobbling motion as if it were no longer fixed in position. Furthermore, stalk formation is inhibited, two histidine kinase signaling proteins fail to localize to the pole, and cell division is defective. GmpA is a cytoplasmic protein that co-localizes with multiple proteins at the cell poles. Biochemical analyses reveal that GmpA self-assembles into filament-shaped polymers that physically interact with several other polar proteins, including the ParB protein at the chromosome origins and two polar hisitinde kinases. Consistent with this, GmpA accumulates into a large complex at the cell pole when overexpressed, creating a structure that is easily distinguishable from normal cytoplasm by cryo-electron tomography. Those proteins that are mislocalized in the absence of GmpA are recruited to this structure, and the distribution of origin regions is also altered to reflect the change in GmpA distribution, suggesting the formation of an enlarged polar domain. We conclude that GmpA is responsible for large-scale organization of the *Caulobacter* cell, acting as a key structural component in the assembly of polar multiprotein complexes and the anchoring of the replication origin.

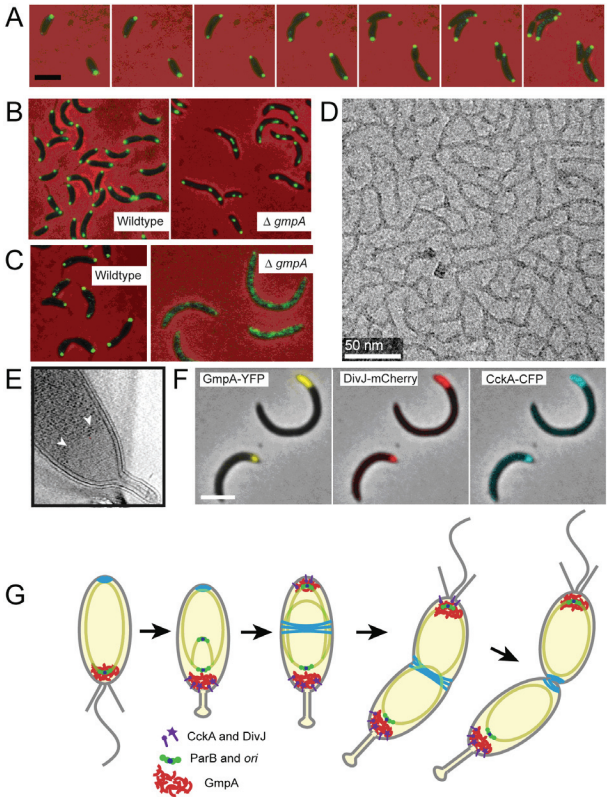


Figure Legend:

(A) GmpA-YFP is localized to the cell poles. Swarmer cells were isolated from a culture of *gmpA-yfp* strain and placed on an agarose pad for microscopic analysis. Overlays of fluorescence (green) and phase contrast images (red shading) at selected times between 0 and 204 minutes are shown.

(B) GmpA anchors the chromosome origin to the cell poles. Here, the origin is marked by CFP-ParB, and its localization is observed in cells expressing GmpA at normal levels (left panel) and Δ*gmpA* cells (right panel) CFP-ParB fluorescence (green) is overlaid on the phase contrast image (red shading).

(C) The CckA histidine kinase is mislocalized in the absence of GmpA. CckA-GFP was observed in the presence (left panel) or absence (right panel) of GmpA expression. The fluorescence signal (green) is overlaid on the phase contrast images (red shading).

* Presenting author

(D) GmpA self-assembles into a structured oligomer. GmpA was expressed and purified from *E.coli* cells, placed on a carbon coated grid, and viewed at 86,000X magnification by transmission electron microscopy.

(E) GmpA overexpression induces the formation of large structures at the cell pole. Here, a 1-voxel thick slice from a tomographic reconstruction of a GmpA overexpressing cell is shown, demonstrating a translucent "plug" at a stalked pole. Arrowheads indicate the transition between the plug and normal cytoplasm.

(F) GmpA recruits transmembrane histidine kinase signaling proteins to the cell poles. CckA-GFP and DivJ-mCherry co-localize with the plug formed by the overexpression of GmpA-YFP. Fluorescent images are overlaid on phase contrast (grey shading).

(G) A model of GmpA function. GmpA (in red) accumulates at the new pole at the swarmer to stalked cell transition, and anchors the newly replicated chromosome origin complex (green and blue) to this location. GmpA also serves as an attachment site for histidine kinases (blue and purple) which are part of the control mechanism for chromosome replication in the next round of cell division.

This work was funded by the DOE GTL project titled "Dynamic spatial organization of multi-protein complexes controlling microbial polar organization, chromosome replication, and cytokinesis". The PI of this multiinstitutional grant is Harley McAdams, Stanford University.

96

GTL

RAPTOR: Robust Alignment and Projection Estimation for TOmographic Reconstruction

Fernando Amat[1]* (famat@stanford.edu), Farshid Moussavi,[1] Kenneth H. Downing,[2] Luis R. Comolli,[2] Mark Horowitz,[1] and **Harley McAdams**[3] (hmcadams@stanford.edu)

[1]Department of Electrical Engineering, Stanford University, Stanford, California; [2]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; and [3]Department of Developmental Biology, Stanford University, Stanford, California

Project Goals: Image enhancement in electron microscopy tomography. Use of statistical imaging and reasoning techniques to improve quality of electron microscopy images. Perform quantitative analysis of the 3D images to acquire biological insights.

We present a method for automatic full-precision alignment of the images in a tomographic tilt series. Full-precision automatic alignment of cryo electron microscopy images has remained a difficult challenge to date, due to the limited electron dose and low image contrast. These facts lead to poor signal to noise ratio (SNR) in the images, which causes automatic feature trackers to generate errors, even with high contrast gold particles as fiducial features. To enable fully automatic alignment for full-precision reconstructions, we frame the problem probabilistically as finding the most likely particle tracks given a set of noisy images, using contextual information to make the solution more robust to the noise in each image. To solve this maximum likelihood problem, we use Markov Random Fields (MRF) to establish the correspondence of features in alignment. The resulting algorithm, called Robust Alignment and Projection Estimation for Tomographic Reconstruction, or RAPTOR, has not needed any manual intervention for the difficult datasets we have tried, and has provided sub-pixel alignment that is as good as the manual approach by an expert user. Our method has been applied to challenging cryo electron tomographic datasets with low SNR from intact bacterial cells, as well as several plastic section and X-ray datasets.

Currently, the software is freely available on the web and several labs (LBL, UCSD, Caltech) are using it for tomographic reconstructions. This work has recently been published in the Journal of Structural Biology in [1].

This work was funded by the DOE GTL project titled "Dynamic spatial organization of multi-protein complexes controlling microbial polar organization, chromosome replication, and cytokinesis". The PI of this multiinstitutional grant is Harley McAdams, Stanford University.

Reference

1. F. Amat and F. Moussavi and L.R. Comolli and G. Elidan and K.H. Downing and M. Horowitz. *Markov random field based automatic image alignment for electron tomography*. Journal of Structural Biology, 2007 (In press).

* Presenting author

97

GTL

Genome-Wide Analysis of Polarly Localized Protein Complexes in *Caulobacter crescentus*: Function, Composition and Spatio-Temporal Dynamics

Edgar Huitema, Sean Pritchard, Sunish Kumar Radhakrishnan, and **Patrick H. Viollier*** (phv1@case.edu)

Department of Molecular Biology and Microbiology, School of Medicine, Case Western Reserve University, Cleveland, Ohio

Project Goals: Identification of polar multiprotein complexes in *Caulobacter crescentus* using a combination of genome-wide genetic and cell biological approaches.

Genetic and cell biological screens were conducted on a genome-wide scale to unearth novel multi-protein complexes that are sequestered to the cell pole in *Caulobacter crescentus*, a prokaryotic model organism for the study of cell polarity. New transposon- and/or plasmid-based tools were developed, furnished with different genetic tags (encoding fluorescence proteins such as mCherry, superfolder GFP or Venus) either for cytological studies or biochemical analysis (affinity purification tags such as TAP), and subsequently used in random genetic screens to isolate (i) strains with translational fusions to polarly localized proteins and (ii) mutants with polar organizational defects. This approach implicated >100 candidates as components of polar multiprotein complexes and/or in controlling the formation of a (sub)complex at the pole. While efforts to define the precise function, composition and spatio-temporal properties of these complexes are still ongoing, two in-depth studies on such protein assemblages localized either at the younger or the older cell pole, were recently completed [1, 2].

Our first study described the role of the TipN scaffolding protein in orchestrating the assembly of the flagellar nanomachine specifically at the younger cell pole [1]. TipN essentially functions as a molecular beacon that marks the correct pole as flagellum assembly site and that recruits building blocks of the flagellum to this pole. TipN is itself already localized to this site when the new pole is formed during cytokinesis. Using cytochemistry, we found TipN to be associated with the cytokinetic machinery that assembles at the division plane and we showed that TipN subsequently remains at the new pole (the flagellum assembly site) that emerges with the completion of cell division, where it directs flagellum construction in the daughter cell. This study provided the first conclusive evidence of an intimate connection between the cell division machinery and the subsequent formation of a multiprotein assemblage at the new cell pole.

Recently, we also reported a multi-protein complex that is specifically associated with the older of the two cell poles [2]. Briefly, we discovered that a developmental protein kinase, DivJ, which regulates the asymmetric division cycle of *Caulobacter* is recruited to the old cell pole by SpmX, a protein homologous to cell wall hydrolytic enzymes. SpmX is required to stimulate DivJ kinase activity and resides in a complex with DivJ, indicating that this class of enzymes, previously believed to have structural roles in cell wall metabolism, can also have profound regulatory functions. SpmX localizes to the old pole prior to DivJ and the enzymatic domain of SpmX is required and sufficient for polar localization. This suggests that organizational signals might exist in the cell wall to nucleate protein assemblages at the poles.

References

1. Huitema E, Pritchard S, Matteson D, Radhakrishnan SK, and Viollier PH. (2006) Bacterial birth scar proteins mark future flagellum assembly site. ***Cell***, 124(5): 1025-37.
2. Radhakrishnan SK, Thanbichler M, and Viollier PH. (2008). The dynamic interplay between a lysozyme homolog and a cell fate determinant drives the asymmetric division cycle of *Caulobacter crescentus*. ***Genes & Development***, 22(2): *in press*.

This work was funded by the DOE GTL project titled "Dynamic spatial organization of multi-protein complexes controlling microbial polar organization, chromosome replication, and cytokinesis". The PI of this multi-institutional grant is Harley McAdams, Stanford University.

* Presenting author

98

GTL

New Methods for Whole-Genome Analysis of Protein Localization

J. Werner, J. Guberman, E. Chen, A. Zippilli, and **Z. Gitai*** (zgitai@princeton.edu)

Department of Molecular Biology, Princeton University, Princeton, New Jersey

Project Goals: To characterize bacterial cell architecture through developing and implementing methods for analyzing the subcellular localization of whole proteomes.

In the past decade, the emerging field of bacterial cell biology has established the importance of properly distributing and organizing subcellular components. It is now clear that virtually every known cellular process requires the precise coordination of the right components to the right cellular address at the right time. To date, however, the labor involved in these studies has required the field to focus on the localization of a small set of previously-characterized proteins. To extend the powerful analysis of protein localization to a whole-genome scale, we have established a series of high-throughput methods for generating, imaging, and analyzing fluorescent protein fusions. Using these new methods, we have determined the localization of over 3,250 proteins in the aquatic bacterium, *Caulobacter crescentus*, and have identified over 450 proteins with distinct non-uniform localizations. We are now in the process of using this data to both deepen our understanding of the functions of specific proteins, as well as to broaden our grasp of the systems-level spatial and temporal organization of a bacterial cell. Our discovery of multiple metabolic pathways with distinct localizations has significant implications for future metabolic engineering efforts. In addition, our high-throughput methods have accelerated our ability to functionally annotate a large number of proteins of previously-unknown function. For example, we have used these methods to identify novel regulators of *Caulobacter* cell division. This poster will discuss the methods we have developed, how they can be adapted to other species of interest, and the systems-level and gene-specific discoveries we have already made.

This work was funded by the DOE GTL project titled "Dynamic spatial organization of multi-protein complexes controlling microbial polar organization, chromosome replication, and cytokinesis". The PI of this multiinstitutional grant is Harley McAdams, Stanford University.

* Presenting author

* Presenting author

Systems Biology Research Strategy and Technology Development

Genomic and Proteomic Strategies

99

GTL

Profiling Microbial Identity and Activity: Novel Applications of NanoSIMS and High Density Microarrays

Eoin Brodie[1]* (elbrodie@lbl.gov), **Jennifer Pett-Ridge**[2] (pettridge2@llnl.gov), Peter Weber,[2] Gary Andersen,[1] Meredith Blackwell,[3] Nhu Nguyen,[4] Katherine Goldfarb,[1] Stephanie Gross,[3] Sung-Oui Suh,[5] James Nardi,[6] Thomas Bruns,[4] and **Paul Hoeprich**[2] (hoeprich2@llnl.gov)

[1]Lawrence Berkeley National Laboratory, Berkeley, California, [2]Lawrence Livermore National Laboratory, Livermore, [3]Louisiana State University, Baton Rouge, Louisiana, [4]University of California, Berkeley, California, [5]American Type Culture Collection, Manassas, Virginia and [6]University of Illinois, Urbana, Illinois

Project Goals: Identification of microorganisms responsible for specific metabolic processes remains a major challenge in environmental microbiology, one that requires the integration of multiple techniques. The goal of this project is to address this challenge by developing a new methodology, "Chip-SIP", combining the power of re-designed oligonucleotide microarrays with nano-scale secondary ion mass spectrometry (NanoSIMS) to link the identity of microbes to their metabolic roles.

Introduction

Identification of microorganisms responsible for specific metabolic processes remains a major challenge in environmental microbiology, one that requires the integration of multiple techniques. The goal of this project is to address this challenge by developing a new methodology, "Chip-SIP", combining the power of re-designed oligonucleotide microarrays with nano-scale secondary ion mass spectrometry (NanoSIMS) to link the identity of microbes to their metabolic roles.

Approach

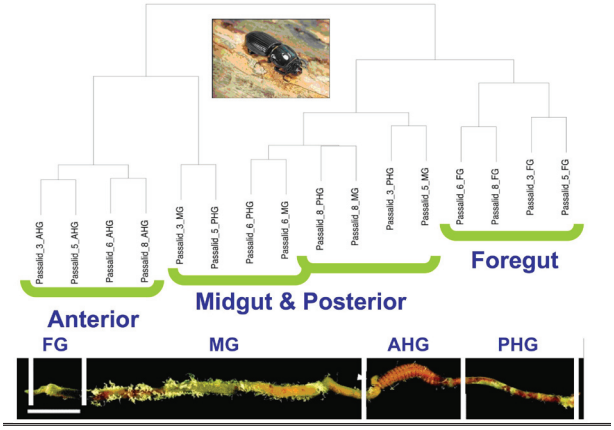
This concept involves labeling of microbial nucleic acids following incubation with a stable isotope-labeled compound (e.g ^{13}C-cellulose or $^{15}N_2$). Extracted RNA is hybridized to a newly engineered high-density oligonucleotide microarray with a conductive surface and higher reproducibility relative to traditional glass/silane microarrays. These advances in array surface chemistry allow successful NanoSIMS analysis of the microarray surface with hybridized nucleic acids, generating isotopic and elemental abundance images of the array surface, and thereby indicating the identity of organisms incorporating the isotopically labeled substrate. To date, we have identified a cyclo-olefin co-polymer plastic (COP) that meets our requirements for these new microarrays (opacity comparable to glass, minimal autofluoresence, adequate hardness and temperature stability to enable surface coating processes). We have coated these COP slides with ~400 angstroms ITO (indium tin oxide) and functionalized the surfaces with alkyl phosphonates. Currently we are testing our ability to manufacture highly reproducible array probe spots using our NimbleGen microarray synthesizer unit to prepare DNA microarrays for eventual analysis by NanoSIMS.

Novel environmental application

Our first environmental application of this approach following validation is to define the microbial biogeography and localize specific metabolic processes within the hindgut of the wood-eating passalid beetle, *Odontotaenius disjunctus*. This microbial community represents a naturally-selected highly-efficient lignocellulose degrading consortium; a thorough understanding of which may aid design and optimization of *in vitro* lignocellulose deconstruction/conversion systems. We have characterized the spatial composition of the bacterial and archaeal communities throughout the beetle gut using our current PhyloChip microarray and have identified distinct population structures within specific hindgut regions. For example, methanogenic archaea were only detected within the anterior hindgut (AHG), the same location which contains the highest relative abundance of fermentative bacteria and anaerobic methane oxidizing archaea.

* Presenting author

We have also optimized DNA/RNA co-extraction and purification, and have standardized donor-mediated direct RNA labeling and microarray hybridization. The relative activity of the microbial species through beetle gut will be determined by direct RNA hybridization. This data will be used to focus probe selection for subsequent synthesis on a NanoSIMS compatible microarray.



Cluster analysis of PhyloChip-based bacterial community composition through the passalid hindgut. Foregut (FG), Midgut (MG), Anterior hindgut (AHG), Posterior hindgut (PHG).

Ongoing work

1. Beetle feeding experiments: At LSU, we are currently cultivating passalid beetles and performing pulse-chase labeling experiments with ^{13}C-labeled glucose. For the feeding experiments, individual *Odontotaenius disjunctus* have been separated into sterile containers and fed a mixture of agar supplemented with 13C-glucose in a 10 ml container. One, 3 and 9 days following isotope label addition, we will harvest several beetles per treatment and perform aseptic dissection of the gut and separation into the four sections. Gut sections are then preserved in RNAlater prior to RNA extraction. In future experiments, additional substrates will be used (^{13}C-cellulose and xylose) and specimens will be contained for 24 hours within a sealed chamber containing air with 99.9 atom% $^{15}N_2$.

2. SIP-Chip method development: Initial NanoSIMS measurements on our newly developed arrays will test our ability to detect ^{13}C in hybridized probe spots. Secondary experiments will be conducted with mixtures of ^{12}C-RNA and ^{13}C-RNA, mixing known amounts of labeled RNA with known amounts of unlabeled RNA, in order to quantify the sensitivity and detection limits of the method. Our initial expectation is that this approach will yield qualitative data (i.e., a spot will be identified as either enriched or not). A third series of tests will be conducted with organisms that have been labeled to differing degrees; creating a standard curve of ^{13}C-RNA samples, with which we can determine our ultimate sensitivity limits and ability to generate quantitative information based upon the degree of isotope incorporation and thus intensity of ^{13}C in individual spots.

100

GTL

NanoSIP: Developing Community Imaging for Phylogenetic and Functional Characterization Using Cyanobacterial Mats

Steven W. Singer,[1]* Jennifer Pett-Ridge,[1] Brad M. Bebout,[2] Tori M. Hoehler,[2] **Leslie E. Prufert-Bebout**[2] (lbebout@mail.arc.nasa.gov), and **Peter K. Weber**[1] (weber21@llnl.gov)

[1]Chemistry, Materials, Earth and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California and [2]NASA Ames Research Center, Moffett Field, California

Project Goals: We have begun working on a new method that will provide correlated oligonucleotide, functional enzyme and metabolic image data to link function and identity. Biomass labeled by incorporation of stable isotope tracers will be combined with oligonucleotide and functional enzyme labels to be imaged by nanometer-scale secondary ion mass spectrometry (NanoSIMS), and referred to as NanoSIP.

Unraveling the metabolic processes of complex microbial communities requires linking the identity of community members with their function. We have begun working on a new method that will provide correlated oligonucleotide, functional enzyme and metabolic image data to link function and identity. Biomass labeled by incorporation of stable isotope tracers will be combined with oligonucleotide and functional enzyme labels to be imaged by nanometer-scale secondary ion mass spectrometry (NanoSIMS), and referred to as NanoSIP. The oligonucleotide and enzyme labels will be elemental labels orthogonal to the stable isotope probes. Preliminary work with a simplified microbial consortium of a filamentous cyanobacterium and a heterotrophic bacterium has allowed us to develop elemental oligonucleotide imaging probes for NanoSIMS based on intracellular fluorine and bromine deposition.

We have begun applying the NanoSIP methodology to hypersaline microbial mats, complex stratified microbial

* Presenting author

communities found in coastal areas. Cyanobacteria of the genera *Microcoleus* and *Lyngbya* are the primary producers in these communities, and they support a diverse assemblage of heterotrophic bacteria. Significant amounts of H_2 are often evolved from these communities under dark, anoxic conditions, and this H_2 evolution has been linked to carbon and nitrogen cycling in the mats. To understand the relationship of H_2 evolution to mat metabolism, we will incubate the mats in the presence of $H^{13}CO_3^-$ and $^{15}N_2$ in "pulse chase" experiments and time course image the consortia by NanoSIMS to determine the fate of C and N. We are currently applying elemental oligonucleotide and enzyme labels to mat sections to link the flow of these stable isotopes to the phylogeny and function.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

101

GTL

NanoSIP: Linking Phylogeny with Metabolic Activity of Single Microbial Cells Using Elemental in Situ Hybridization and High Resolution Secondary Ion Mass Spectrometry

Sebastian Behrens,[1] Tina Lösekann,[2] Jennifer Pett-Ridge,[3] Wing-On Ng,[2] Bradley S. Stevenson,[4] David A. Relman,[2,5] Alfred M. Spormann,[1] and **Peter K. Weber**[3]* (weber21@llnl.gov)

[1]Department of Chemical Engineering and Civil and Environmental, Stanford University, Stanford, California; [2]Departments of Microbiology and Immunology, and of Medicine, Stanford University, Stanford, California; [3]Chemistry, Materials, Earth and Life Science Directorate, Lawrence Livermore National Laboratory, Livermore, California; [4]Department of Botany and Microbiology, University of Oklahoma, Norman, Oklahoma; and [5]Veterans Administration Palo Alto Health Care System, Palo Alto, California

Project Goals: We are developing a new technique-NanoSIP-to determine nutrient uptake and translation at the single cell level. The method combines in situ phylogenic and immuno labeling methods with stable isotope probing (SIP) and nanometer-scale secondary ion mass spectrometry (NanoSIMS) analysis to enable microbial identity and function to be probed in intact samples.

Linking phylogenetic and functional information in complex communities is a key challenge in microbial ecology. To address this need, we are developing a new technique—NanoSIP—to study nutrient uptake and translation at the single cell level. The method combines *in situ* phylogenic and immuno labeling methods with stable isotope probing (SIP) and nanometer-scale secondary ion mass spectrometry (NanoSIMS) analysis to enable microbial identity and function to be probed in intact samples. In this study, we demonstrate NanoSIP using a double labeling method, in which elemental labels are combined with fluorescent *in situ* hybridization methods (EL/FISH) to target 16S rRNA to enable simultaneous visualization of identity and function during NanoSIMS analysis. Correlated fluorescence and NanoSIMS imaging is also achieved in simple samples.

We developed the EL/FISH method with monocultures and simple mixtures of *E. coli*, and *Rhizobium* species. A general bacterial probe was used to label all species, and an α-proteobacteria probe was used to target *Rhizobium*. Fluorine and bromine elemental labels were used to enable elemental imaging of labeled cells by NanoSIMS. We overcame F and Br background levels in the samples by enhancing the elemental labels with CARD-FISH methods (catalyzed reporter deposition-fluorescent *in situ* hybridization) that labeled the target microorganisms with both a fluorophore and an elemental label (F or Br). Elemental label in the targeted microorganisms exceeded background levels, enabling the target organisms to be readily distinguished from non-target microorganisms. Specificity was confirmed by performing the CARD-EL/FISH procedure with an oligonucleotide probe with a non-matching sequence (nonsense probe).

We then applied CARD-EL/FISH as part of a NanoSIP study of a microbial consortium that consists of a heterotrophic *Rhizobium* α-proteobacterial epibiont and a filamentous *Anabaena* cyanobacterium (Figure A). The epibionts are believed to receive N from *Anabaena* because they attach to the *Anabaena* heterocysts, which are specialized nitrogen fixing cells. To test this hypothesis, the two species were incubated independently and in co-culture in $H^{13}CO_3^-$ and $^{15}N_2$. The α-proteobacteria probe was then used with CARD-EL/FISH to label the epibiont with F. Fluorescence was first imaged, and then the imaged locations were analyzed by NanoSIMS to show the cells with the F label and the distribution of newly fixed ^{13}C and ^{15}N (Figure B-D). In this way, phylogenetic information and functional activities are determined in the same analysis. In the monocultures, *Anabaena* fixed C and N, and the epibiont did not. In co-culture, the attached epibiont acquire both the ^{13}C and ^{15}N label, demonstrating that the epibiont gain both

* Presenting author

nutrients from the *Anabaena*. Note that in the given example, the color scaling for the ^{15}N enrichment image makes the epibiont appear only weakly enriched when it is in fact significantly enriched ($\delta^{15}N$ ~ 1000 parts per thousand relative or 2 times natural abundance). The color scaling is set by the *Anabaena* vegetative cells, which are very enriched in ^{15}N ($\delta^{15}N$ ~ 13,000 parts per thousand, which is 14 times higher than natural abundance). As previously observed, mature heterocysts, which fix nitrogen but do not divide, are less enriched in ^{15}N than the vegetative cells because their need for newly fixed nitrogen is low [1-3]. For the same reason, mature heterocysts are only slightly enriched in ^{13}C.

$$\delta^{15}N = [(^{15}N/^{14}N)_{unknown}/(^{15}N/^{14}N)_{standard} - 1] \times 1000;$$
$$\delta^{13}C = [(^{13}C/^{12}C)_{unknown}/(^{13}C/^{12}C)_{standard} - 1] \times 1000$$

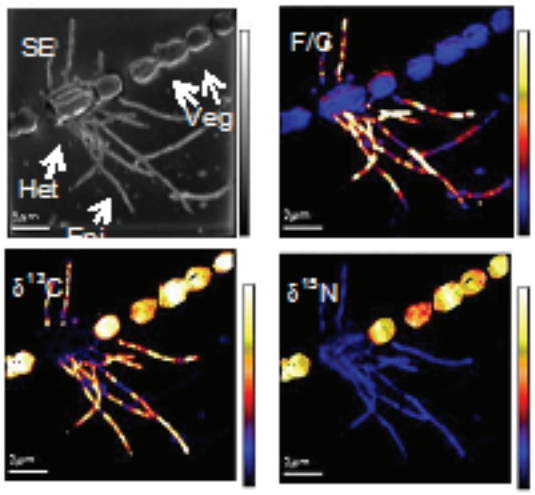


Figure. NanoSIMS images of a chain of 8 cells from the filamentous cyanobacterium *Anabaena* sp. SSM-00 with cells of the α-Proteobacterium epibiont *Rhizobium* sp. WH2K attached to a heterocyst. Images were taken after 24 h of incubation with $H^{13}CO_3^-$ and $^{15}N_2$. SE: Secondary electron image showing the location of all cells. F/C: Flourine is enriched in the epibiont relative to *Anabaena* after *in situ* hybridization with an α-Proteobacteria specific probe (scale, 0 – 0.2). $\delta^{13}C$: ^{13}C enrichment image showing the relative uptake of newly fixed carbon (scale, 0 – 300 parts per thousand). $\delta^{15}N$: ^{15}N enrichment image showing the relative uptake of newly fixed nitrogen (scale 0 – 15,000 parts per thousand). The color bars indicate the relative enrichment of the isotope in the image. Het, heterocyst; Veg, vegetative cells; Epi, epibiont.

References

1. Wolk, CP, SM Austin, J Bortins and A Galonsky 1974. Autoradiographic localization of ^{13}N after fixation of ^{13}N-labeled nitrogen gas by a heterocyst-forming blue-green alga *The Journal of Cell Biology* **61**: 440-453.
2. Wolk, CP, J Thomas, PW Shaffer, SM Austin and A Galonsky 1976. Pathway of nitrogen metabolism after fixation of ^{13}N-labeled nitrogen gas by the cyanobacterium *Anabaena cylindrica J. Biol. Chem.* **251**: 5027-5034.
3. R. Popa, P.K. Weber, J. Pett-Ridge, J.A. Finzi, S.J. Fallon, I.D. Hutcheon, K.H. Nealson and D.G. Capone (2007) Carbon and nitrogen fixation and metabolite exchange in and between individual cells of *Anabaena oscillarioides*, *The International Society of Microbial Ecology (ISME) Journal* 1: 354-360.

102

GTL

High Throughput Comprehensive and Quantitative Microbial and Community Proteomics

Gordon A. Anderson, David J. Anderson, Kenneth J. Auberry, Mikhail E. Belov, Stephen J. Callister, Therese R.W. Clauss, Jim K. Fredrickson, Xuixia Du, Kim K. Hixson, Navdeep Jaitly, Gary R. Kiebel, Mary S. Lipton, Eric A. Livesay, Anoop Mayampurath, Matthew E. Monroe, Ronald J. Moore, Heather M. Mottaz, Carrie D. Nicora, 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, Karl Weitz, Rui Zhang, Rui Zhao, and **Richard D. Smith*** (rds@pnl.gov)

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

Project Goals: The goal of this project is to develop and apply a capability for detecting and identifying large numbers of proteins from microbial proteomes, and for improving the understanding of complex microbial communities.

Significance: Capabilities for quantitative proteomics measurements have been developed that can now achieve high levels of throughput and quality, allow broad studies of e.g. diverse microbial systems and communities, and provide new systems level biological insights.

With recent advances in whole genome sequencing for a growing number of organisms, biological research is increasingly incorporating higher-level systems perspectives and approaches. Advancing the understanding of microbial and bioenergy related systems is at the heart of DOE's Genomics:GTL program, and present immense

* Presenting author

challenges. For example, microbial cells in nature rarely exist as individual colonies, but interact with other neighboring microbes and with their environment, thus creating an ecosystem. The challenges associated with studying such complex systems are significant due to the inherent biological complexity and number of possible interactions, and to the limitations in current technologies that will need to be addressed to allow us to more completely characterize these complex systems.

A key aspect for acquiring such biological understandings is the ability to quantitatively measure the array of proteins (i.e., the proteome) for the system being studied under many different conditions. Ultimately, such measurements and the resulting insight into biochemical processes can enable development of predictive computational models that could profoundly affect environmental clean-up, climate-related understandings, and energy production e.g. by providing understandings of energy production-related activities on the environment.

First among the basic challenges associated with making useful comprehensive proteomic measurements is identifying and quantifying large sets of proteins whose relative abundances typically span many orders of magnitude. Additionally, these proteins may vary broadly in chemical and physical properties, have transient and low levels of modifications, and can be subject to endogenous proteolytic processing.

A second key challenge is in making proteomics measurements of sufficiently high throughput so as to truly enable systems biology approaches. A major limitation to date has been that obtaining higher throughput has required significant trade-offs in both the quality of measurements and the "depth" of proteome coverage. Higher throughput measurement capabilities that can also provide broad proteome coverage can allow fundamentally different approaches to be taken in experiment design, and enable measurements that provide practical insights into issues such as biological variation. High throughput measurements also allow the application of proteomics to a much larger array of organisms. Recent efforts e.g. have demonstrated that it is practical to conduct proteomics measurements at a modest cost in parallel with genome sequencing efforts and in a manner that can also augment annotation efforts.

A third challenge is related to the sensitivity achieved in proteome measurements. Higher sensitivity measurements are often needed to make practical high throughput measurements, as well as the use of stable isotope labeling, methods for sub-cellular fractionation, etc.

The proteomics program at PNNL is addressing these issues in the context of a range of biological applications (see presentation by Mary Lipton, et al.), and in collaboration with a number of Genomics:GTL researchers. An extensive high throughput proteomics pipeline has been developed and steadily refined that is based upon the application of ultra-high performance separation-mass spectrometry approaches. These measurement capabilities are integrated with an advanced informatics pipeline that provides the essential tools needed to deal with the high data production rates. The Accurate Mass and Time (AMT) tag approach developed at PNNL has proven essential for enabling both effective quantitation and the desired throughput (needed to also access the effectiveness of quantitation). The PNNL program has generated the largest quantitative proteomics datasets to date in conjunction with GTL collaborators, and has effectively applied methods such as subcellular fractionation in conjunction with proteomics measurements to further extend the biological insights achieved based upon protein localization. On the basis of the increased throughput being achieved, we believe that proteomics can now at modest cost provide an important adjunct to *all* genome annotation efforts.

This presentation will summarize the present state of proteomics measurement capabilities and describe efforts in progress that are significantly extending or improving throughput, coverage, sensitivity, and quality of quantitation, and why these improvements are important. Advances in the proteome measurement technology based upon new mass spectrometry instrumentation and approaches will be described. In addition the crucial role of the data processing informatics pipeline that has been developed to provide the necessary throughput will be discussed, as well as developments that are providing more effective protein identifications and improved coverage of protein modifications. The presentation will also summarize remaining challenges related to throughput and measurement quality, the large opportunities that can be derived from more and better measurements, and a description of technology and informatics advances that are expected to address these needs.

The presentation will conclude with a discussion of how these proteomics capabilities are expected to advance systems biology approaches, and specifically the interests of the Genomics:GTL program, and how the technical capabilities developed for proteomics can be applied to further augment systems biology by enabling more effective metabolomics measurements.

Acknowledgements: This research is supported by the Office of Biological and Environmental Research of the U.S. Department

* Presenting author

of Energy. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute through Contract No. DE-AC05-76RLO 1830.

103

GTL

Proteomics Driven Analysis of Microbes and Microbial Communities

Joshua Turse,[1] Stephen J. Callister,[1] Margaret F. Romine,[1] Kim K. Hixson,[2] Samuel O. Purvine,[2] Angela D. Norbeck,[1] Matthew E. Monroe,[1] Xuixia Du,[1] Feng Yang,[1] Brain M. Ham,[1] Carrie D. Nicora,[1] Richard D. Smith,[1] Jim K. Fredrickson,[1] and **Mary S. Lipton**[1]* (mary.lipton@pnl.gov)

[1]Biological Sciences Division and [2]Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington

Project Goals: Key to understanding biological systems in support of U.S. Department of Energy's (DOE) missions and science is the ability to quantitatively measure the array of proteins, also termed the proteome, in plants, prokaryotic cells and communities of cells. The ability to make global measurements of gene and protein expression with the desired comprehensive qualities is now feasible through new, advanced technologies. The present project has involved the application of PNNL proteomics capabilities to a range of biological system, with an initial focus on microbial systems, and including more recent extension to simple microbial communities. Our plan is to continue the collaborative application of proteomics measurement capabilities to microbial systems, and to extend such efforts to the proteomic characterization of microbial communities and plant systems. We will continue studies of systems including *Shewanella oneidensis* MR-1 and related species, *Rhodobacter sphaeroides* 2.4.1, *Geobacter sp.*, *Caulobacter crescentus*, and *Pelagibacter ubique*. We will also extend our initial work with microbial communities, to communities in the Columbia River and the open sea. Additionally, we will extend our collaborative applications to include plant systems e.g. poplar in studies associated with bioenergy applications.

Collaborators:
Phil Hugenholtz/Falk Warneke (*JGI*)
Lucy Shapiro *(Stanford University)*
Steve Giovannoni (*Oregon State University*)
Derek Lovley *(University of Massachusetts)*
Tim Donohue (*University of Wisconsin-Madison)*
Sam Kaplan *(UT-Houston Medical School)*
Jim Teidje *(Michigan State University)*
Shewanella Federation

Significance: Characterization of microbiological systems using comprehensive global proteomic studies enhances scientific understanding through improved annotation of genomic sequences, elucidation of phenotypic relationships between environmentally important microorganisms, characterization of the metabolic activities within microbial communities, and identification of post-translationally modified proteins.

Proteomic applications support DOE missions and science by exploiting microbial function for purposes of bioremediation, energy production, and carbon sequestration among other important areas. Inherent to exploiting microbial function is the ability to rapidly acquire global measurements of the proteome (i.e., the proteins expressed in the cell). This applications project exploits the proteomics pipeline at PNNL to address organism-specific scientific objectives developed in conjunction with biological experts for a number of different microbes. In our poster, we highlight biological results to date for investigations of *Shewanella* species, *Caulobacter crescentus*, *Peligabacter ubique, C. crescentus*.

The proteome can play an integral role in the protein annotation of sequenced genomes by cross validating expressed proteins with genome sequences predicted to encode proteins. For example, proteomics can be used to cross validate genome annotations by verifying that predicted genes do in fact encode proteins, resolve conflicts between different gene prediction algorithms, identify erroneous gene termini predictions, provide evidence for frameshift events that lead to alternative protein products, and provide evidence for intergenic region gene products missed in the gene calling process. Many of these omissions have been observed in initial genome-to-proteome investigations performed on *Caulobacter crescentus*, *Shewanella oneidensis* MR-1, *Shewanella baltica* OS185 and *Shewanella baltica* OS195 through the use of stop-to-stop translation of recently published genomes for these organisms. Using *Shewanella baltica* OS195, we clearly show the usefulness of proteomics in the annotation process of a microbe.

Proteome comparisons across multiple microorganisms can play an integral role in defining phenotypic similarities and differences. We highlight a specific study of 11 *Shewanella* species that have sequenced genomes in which we demonstrated the ability to characterize the proteome of an organism using the genome sequence of a closely related organism. This demonstration has

* Presenting author

broad implication for conducting proteomic research on environmentally important organisms that do not have sequenced genomes. Additionally, we calculated a degree of phenotypic relatedness (phyloproteomic relatedness) from the 11 proteomes. We also highlight another study of 17 diverse microorganisms that evaluated the concept of the core genome – a set of orthologous genes commonly derived in bacterial genomic studies, on the proteome level of the proteome. Proteomics revealed a core proteome of potentially essential proteins to bacterial life, and also revealed unique lifestyle differences that are dependent on culture environment.

Our proteomic capabilities have been applied to characterize both the open ocean community in relation to *Peligabacter ubique* and the microbial community isolated from the termite (*Nasutitermes corniger*) hindgut. The proteome characterization of these microbial communities presents a challenging application, and we are in the early stages of seeking to understand the ecology of these communities at the protein expression level and how this protein expression relates to the interaction of microbe with the environment and within the community. We show for *P. ubique* how significant expression of the proteins involved in transport of metabolites and metals are indiciative of the environmental metabolic requirements of this organism.

Proteins regulate their function through expression levels and post-translational modifications, which can both be measured by proteomic analyses. Focusing on the characterization of the cell cycle in *C. crescentus*, we examined growth under carbon and nitrogen limitation conditions along with temporal resolution time course samples to provide new insights on how this organism responds to its environment through genomic, proteomic, and ultimately morphologic strategies. We present results from the characterization of phosphorylation patterns of this organism, which revealed phosphorylation sites at threonine, serine, tyrosine and aspartate. Additionally, nine proteins observed to be up regulated through these modifications are likely involved in elevated signaling processes associated with an adaptive response to the carbon starved growth environment.

Additional information and supplementary material can be found at the PNNL proteomics website at http://ober-proteomics.pnl.gov/.

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. DE-AC05-76RLO 1830. Parts of this project is component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus.

104

GTL

Biofilm Growth Technologies for Systems Biology

Jeff S. McLean* (jmclean@jcvi.org)

J. Craig Venter Institute, La Jolla, California

Microbial biofilms possess spatially and temporally varying metabolite concentration profiles at the macroscopic and microscopic scales. This results in varying growth environments that may ultimately drive species diversity, determine biofilm structure and the spatial distribution of the community members. Much work has been done to understand biofilm development processes however; challenges arise when applying high throughput systems biology technologies such as transcriptomics and proteomics since these are bulk techniques and capture an average of the population. Dealing with biofilm complexity is a major challenge. Controlled cultivation techniques for biofilm growth can help reach a steady state for various stages of biofilm formation which includes attachment, monolayer formation, mature biofilm formation and detachment processes.

Communities of bacteria in nature that are attached to surfaces exhibit a high degree of complexity in terms of species composition, structure, and spatial distribution of cellular functions. It is generally accepted that in most biofilms (single or multi-species), metabolite concentration gradients develop as a consequence of diffusion limitations and cellular metabolism. Metabolite concentration fluxes may also vary widely with hydrodynamic flow since this will impact the boundary layer thickness and serve to vary the availability of nutrients and the removal of byproducts. Under hydrodynamic conditions, the residence time of small metabolites such as organic acids and quorum sensing molecules may cause local changes in gene expression and cell metabolism which lead to changes in biofilm architecture and the underlying substratum (e.g., erosion/corrosion). Metabolic byproduct accumulation for example is known to have regulatory effects in planktonic cell populations and therefore is likely to play a major role in gene regulation inside a diffusion-limited biofilm. There may also be a direct relationship between metabolite concentrations and the architecture of the biofilm of respiratory bacteria due to limitations imposed by the electron donor and/or acceptor availability. Global regulatory triggers for observed

* Presenting author

coordinated behavior such as complex tower formation and swarming dispersal may therefore be a result of localized nutrient limitation or metabolite buildup in diffusion-limited regions of the biofilm. Overall, these spatial variations in metabolite concentrations and hence their fluxes through biofilms are of fundamental importance to biofilm structure and function. It is these micro- and macro- scale gradients that likely control the development, spatial organization and sustainability of mixed species microbial communities.

Overall, understanding and controlling the environmental conditions and biofilm stage of growth allows better interpretation of bulk omics techniques. Standardization of biofilm growth techniques is in development however one technique may not be adequate to generate samples which are compatible for all applications. Ultimately as sample size requirements for high throughput techniques are lowered, spatial resolution of gene and protein expression within biofilm colonies become feasible when coupled to sample extraction techniques. In addition, the development and application of new technologies which allows capture of information non-invasively is critical to obtain the macro and microscale metabolic and phenotypic status of the cells that are to be sampled.

105

GTL

Experimental Proteogenomics Approaches to Investigate Strain Variations and Molecular Level Activities of a Natural Microbial Community

Robert L. Hettich[1]* (hettichrl@ornl.gov), Nathan VerBerkmoes,[1] Paul Abraham,[1] Chongle Pan,[1] Brian Erickson,[1] Manesh Shah,[1] Doug Hyatt,[1] Denise Schmoyer,[1] Vincent Denef,[2] Paul Wilmes,[2] Ryan Muller,[2] Steve Singer,[3] Michael Thelen,[3] and **Jillian Banfield**[2] (jbanfield@berkeley.edu)

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

Project Goals: Experimental MS-based proteomics technologies and bioinformatic approaches are being developed to characterize microbial communities in their natural settings. This is vital for the accurate elucidation of how these consortia adapt and respond to their environment. Recent work has focused on a large-scale study of temporally and/or spatially resolved AMD biofilm samples in an effort to probe the genomic diversity of the AMD system, and to evaluate how proteomic information can be obtained on samples somewhat distant from the genomic sequencing. Due to the staggering amount of proteogenomics data for the AMD system, we have carefully designed and populated a MySQL database that captures all of the AMD measurements to date. This organizes the large volume of data into fields that can easily be interrogated by various query tools.

While many different microbial species can be grown as isolates and studied in the laboratory, their behavior in natural environmental communities can be significantly different, since they have to cooperate and compete for natural resources. To this end, the development of experimental and computational approaches to characterize microbial communities in their natural settings is vital for the accurate elucidation of how these consortia adapt and respond to their environment. Because the complexity of many natural microbial communities exceeds the current measurement capabilities of analytical techniques, it is advantageous to start with a low complexity environmental microbial consortia. In this respect, the acid mine drainage (AMD) microbial system is ideal. Sufficient biomass is readily accessible to enable molecular level evaluation by a variety of genomic, proteomic, and biochemical techniques. This permits coordination of different analytical measurements on the same samples, thereby providing the ability to integrate the datasets for extraction of biological information.

Whole community genomics serves as the underlying core for almost all of the subsequent measurements and evaluations of microbial consortia. The depth and quality of genome annotation, including information about strain variation, is critical for the ensuing proteomic and biochemical measurements. Initial genome work in the Banfield lab (UC-Berkeley) resulted in good coverage of the most abundant bacteria (*Leptospirillium* groups II and III) and archaea (*Ferroplasma*). This enabled a fairly deep proteome measurement of the most abundant species; however, many abundant peptides measured in the proteome samples could not be matched to anything in the genome database. Recent work has greatly expanded the genome annotation of both bacteria and archaeal species in the AMD samples (see UCB poster), providing a much richer database from which to mine proteome data. The presence of at least five more archaeal species, several low abundance bacterial species, along with three novel nanoarchaea in the updated genomic dataset now serves as the basis for deeper proteome coverage, enabling more comprehensive insights. Even with fairly limited strain

* Presenting author

variation in the AMD system, there are daunting challenges for the proteomic analysis. In particular, much of our previous work has focused on differentiating unique and non-unique peptides, in order to map them to specific organisms or strains. The presence of closely related microbes prompts the need for bioinformatic tools to deal with semi-unique peptides. Computation algorithms are under development to determine the optimum way to classify both non-unique and semi-unique peptides, so as figure out how assign them to the appropriate proteins for more accurate quantification determinations.

In conjunction with the Banfield UCB research group, a fairly large scale study is underway to characterize spatially and temporally resolved AMD samples. The goal of this work is two-fold; to probe the genomic diversity of the AMD system, and to evaluate how proteomic information can be obtained on samples somewhat distant from the genomic sequencing. We are terming the latter aspect as peptide-inferred genome typing, or PIGT (described in the 2007 UCB poster). We have completed about 30 full-scale PIGT proteome measurements (with replicates) of spatially and/or temporally resolved AMD samples. These were done without extensive sample fractionation into soluble and membrane segments; rather, the goal was a faster screen for moderately deep proteome coverage on as many samples as possible. Results indicated that there are only two major stain types (along with a recombinant version) of the abundant *Leptospirillium* group II bacteria across all the locations. FISH imaging is used to help characterize the distribution of organism types, thereby assisting in the deciphering of the mass spectrometric data (Denef, *et. al*. in prep).

In conjunction with the Thelen LLNL research group, recent work has also focused on deeper characterization of the extracellular fraction of the AMD sample. Initial work indicated the presence of abundant, unique cytochromes in this fraction; subsequent work is directed at a deeper examination of the range of other important extracellular proteins. For this work, various chromatographic methodologies are used to fractionate intact proteins for eventual MS characterization. In particular, work is focused on the purification and characterization of abundant unknown proteins from the extracellular medium. An integrated top-down, bottom-up MS approach is being used to characterize the resulting proteins. This approach provides information about the degree of post-translational modifications, in particular signal peptide cleavages, for the representative proteins.

To date, a staggering amount of MS proteome data has been acquired for the AMD samples, and is becoming almost unmanageable in terms of extracting information. To this end, we have carefully designed and populated a MySQL database that captures all of the AMD measurements to date. This organizes the large volume of data into fields that can easily be interrogated by various query tools. Much effort has gone into designing and populating the database in such a way that not only direct collaborators but also the general scientific community can easily query it in a variety of ways. Initial query tools were designed to extract information about the presence of specific proteins across all of the samples, as well as the most abundant redundant proteins in all samples. This will greatly aid in correlating and comparing extensive datasets to extract biological information that might provide a detailed insight into the functional activities of natural microbial communities. Since proteomics data is archived and can be re-searched, we are currently re-mining all of the major proteome datasets (6 different proteomes with extensive fractionation and analyses time) against the new genomics databases containing new archaeal species, several low abundance bacterial species, and novel nanoarchaea as well as viral sequences. The new genomic databases should provide new insight into the proteomics data and thus the structure and physiology of the AMD biofilms; note that we were virtually blind to this level of detail with original limited genome databases.

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

106

GTL

Proteogenomics of Two Environmentally Relevant Microbial Communities

Paul Wilmes[1]* (pwilmes@berkeley.edu), Gregory J. Dick[1]* (gdick@berkeley.edu), Anders F. Andersson,[1] Mark G. Lefsrud,[3] Margaret Wexler,[4] Manesh Shah,[5] Robert L. Hettich,[6] Michael P. Thelen,[7] Philip L. Bond,[8] Nathan C. VerBerkmoes,[6] and **Jillian F. Banfield**[1,2]

[1]Department of Earth and Planetary Sciences and [2]Department of Environmental Science, Policy, and Management, University of California, Berkeley, California; [3]McGill University, Bioresource Engineering, Ste-Anne-de-Bellevue, Quebec, Canada; [4]School of Biological Sciences, University of East Anglia, Norwich, United Kingdom; [5]Life Sciences Division and [6]Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee;

* Presenting author

[7]Biosciences Directorate, Lawrence Livermore National Laboratory, Livermore, California; and [8]Advanced Wastewater Management Center, University of Queensland, St. Lucia, Queensland, Australia

Project Goals: The aim of this project was to apply genomic and proteomic techniques to two distinct microbial ecosystems (acid mine drainage biofilms and activated sludge), and to develop novel means of assigning the resulting information to functionally relevant organismal groups within both systems.

Community genomic and proteomic methods have demonstrated their ability to provide unprecedented insight into microbial ecosystems of fundamental environmental importance. However, due to the complexity of these systems, numerous challenges arise with regards to assignment of genomic and proteomic information to specific organisms within the analyzed communities. Here we discuss the application of genomic and proteomic techniques to two distinct microbial ecosystems differing substantially in complexity, and describe novel ways of assigning both genomic and proteomic information to functionally relevant organismal groups.

Acid mine drainage (AMD) is a worldwide environmental problem that is driven in part by microorganisms that catalyze pyrite (FeS_2) dissolution. In the Richmond Mine at Iron Mountain, CA, limited energy sources and extreme conditions (low pH, high concentrations of toxic metals) restrict microbial community diversity to only a handful of dominant organisms. From these relatively simple communities, genomic DNA sequence has been used to reconstruct near-complete genomes for two bacteria (*Leptospirillum* groups II and III), five Archaea from within the *Thermoplasmatales* (*Ferroplasma*, A-plasma, E-plasma, G-plasma, and I-plasma), and several novel Archaeal lineages (ARMAN 2, 4, and 5). A major challenge of sequence-based community genomics – particularly shotgun sequencing of small fragments – is "binning", or assignment of genomic fragments to their host genomes. For the dominant members of the AMD ecosystem we were able to employ perhaps the most effective means of binning: assembly of genomic sequence into large deeply-sampled fragments that include phylogenetically informative markers such as rRNA genes. However, most natural microbial communities harbor tremendous diversity at both the species and genomic level that (on the sequencing scales employed to date) have precluded significant assembly. Further, the binning problem is particularly acute for low-abundance organisms that have been shallowly sampled, even in our low diversity ecosystem. The result is community genomic datasets where only a small portion of fragments contain phylogenetically-informative genes, leaving a large number of anonymous fragments. The deeply-sampled, manually curated genomes from the Iron Mountain AMD system essentially provide an answer key with which the performance of binning methods can be evaluated. Genomic signatures such as oligonucleotide frequencies have previously been shown to be distinct among organismal genomes and are an attractive option for binning of metagenomic data because they require no prior knowledge of the sample in question and are thus not susceptible to biases of the current sequence databases. We used tetranucleotide frequency and self-organizing maps (SOMs) to evaluate a dataset of AMD community genomic sequence that included both previously assembled/identified sequences as well as unassigned sequence fragments. The tetra-SOM effectively resolved most of the assembled genomic sequence and revealed previously unrecognized regions of tetranucleotide frequency signature that correspond to novel low-abundance organisms and putatively extrachromosomal elements of dominant organisms (i.e. phage or plasmid). The ability to resolve genomes was a function of phylogenetic relatedness rather than G+C content: distantly related genomes with identical G+C content were effectively resolved whereas more closely related genomes with distinct G+C content showed some region of overlap. Overall, our results demonstrate that tetra-SOM is a valuable method of binning and visualizing community genomic data.

Biological wastewater treatment plants are operated throughout the world and harbor extensive microbial diversity. Activated sludge undergoing alternating anaerobic and aerobic regimes is enriched for specific polyphosphate accumulating organisms (PAOs) that enable enhanced biological phosphorus removal (EBPR) from wastewater. Dominant PAOs in such systems have so far eluded cultivation attempts and, hence, have only been putatively named "*Candidatus Accumulibacter phosphatis*" (*A. phosphatis*) based on molecular studies. With the recent availability of extensive metagenomic sequences from *A. phosphatis*-dominated sludges, we were able to employ high-resolution community proteomics to identify key metabolic pathways in *A. phosphatis*-mediated EBPR. Furthermore, we evaluated the contributions of co-existing strains within the dominant population. Results highlight the importance of denitrification, fatty acid cycling and the glyoxylate bypass in EBPR. Despite overall strong similarity in protein profiles under anaerobic and aerobic conditions, fatty acid degradation proteins were more abundant during the anaerobic phase. Using tetra-SOM, we uncovered that a large fraction of previously unclassified scaffold fragments cluster with *A. phosphatis.* Importantly, proteins encoded by these

* Presenting author

scaffolds were identified by proteomics. Hence, these previously unassigned genomic fragments probably have functional significance within EBPR. By comprehensive genome-wide alignment of orthologous proteins, we uncovered strong functional partitioning for enzyme variants involved in both core-metabolism and EBPR-specific pathways among the dominant strains. These findings emphasize the importance of genetic diversity in maintaining the stable performance of EBPR systems and demonstrate the power of integrated cultivation-independent genomics and proteomics for analysis of complex biotechnological systems.

This research was supported by a grant from the DOE Genomics:GTL program (2005).

107

GTL

Structure and Function for Novel Proteins from an Extremophilic Iron Oxidizing Community

Korin Wheeler,[1]* Yongqin Jiao,[1] Steven Singer,[1] Adam Zemla,[1] Nathan VerBerkmoes,[2] Robert Hettich,[2] Daniela Goltsman,[3] Jill Banfield,[3] and **Michael Thelen**[1] (thelen1@llnl.gov)

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

Project Goals: With integrated metagenomic and proteomic datasets as a foundation, we are establishing a combination of computational and experimental methods to determine the structure and function of the several hundred proteins of unknown function within our well-defined, acid mine drainage model system. Our studies will enable a system-wide understanding of each protien and its role in cellular pathways and intracellular communication in this extremophilic microbial community.

As information from proteomic and genomic analyses rapidly escalates, the number of genes and proteins of unknown function continually expands. Yet methods to understanding these novel proteins, often key to understanding unique aspects of niche adaptation, are only just emerging. Because extreme environments are geochemically distinct and biologically limiting, low complexity ecosystems like that of acid mine drainage are ideal for such studies. Genomic and proteomic analysis of samples collected from the Richmond Mine in Iron Mountain (Redding, CA) have provided an initial survey of the genes and proteins that function within the community; however, it remains unclear how the numerous unique proteins facilitate survival under these conditions. With integrated metagenomic and proteomic datasets as a foundation[2], we are using a combination of computational and experimental methods to determine the structure and function of the several hundred proteins of unknown function within our model system.

Initial studies center on a high-throughput computational approach for predicting structure and function for 421 novel proteins from the dominant species in the community. We have developed a structural modeling system to compare these proteins to those of known structure (AS2TS)[2], resulting in the assignment of structures to 360 proteins (85%) and functional information for up to 75% of the modeled proteins. Detailed examination of the modeling results reveals the roles of many of the novel proteins within the microbial community. Protein classes (e.g., hydrolases, oxidoreductases) and families (metalloproteins, tetratricopeptide [TPR] repeats) that are highly represented in the community are now being targeted in experimental work. Complementing structure-function studies are biochemical and molecular biological techniques. Affinity chromatography has enabled enrichment of novel proteins with specific functions or active sites moieties. Environmental DNA clone libraries have facilitated screening of bacterial colonies for hydrolytic enzymes, including proteases, phosphatases, amylases, and lipases. Further to this molecular approach, a bacterial two-hybrid screen has been established to identify proteins interacting with our novel proteins, such as a novel iron oxidizing cytochrome and 27-repeat TPR protein.

References

1. Ram et al, 2005, *Science* 308:1915-20, "Community Proteomics of a Natural Microbial Biofilm."
2. Zemla et al, 2005, *Nucleic Acids Res* 33 (Web Server issue):W111-5, "AS2TS system for protein structure modeling and analysis."

This work was funded by the DOE Genomics:GTL Program and was performed under the auspices of the DOE by the University of California, Lawrence Livermore National Laboratory under contract W-7405-Eng-48.

* Presenting author

108

GTL

Purification and Characterization of Viruses From an Acid Mine Drainage System

Kimberly Pause[1]* (kpause@marine.usf.edu), Christine Sun,[2] Paul Wilmes,[2] Brett Baker,[2] Luis R. Comolli,[3] Chongle Pan,[4] Robert Hettich,[4] Nathan VerBerkmoes,[4] Jillian F. Banfield,[2] and **Mya Breitbart**[1]

[1]University of South Florida, Saint Petersburg, Florida; [2]University of California, Berkeley, California; [3]Lawrence Berkeley National Laboratory, Berkeley, California; and [4]Oak Ridge National Laboratory, Oak Ridge, Tennessee

Project Goals (Abstracts 109-111): The objective of this project is to develop integrated genomic and proteomic (proteogenomic) methods to study virus-microbial community interactions in bioreactor-grown and natural bacterial/archaeal biofilm communities. Simultaneous genomic analysis of microbes and viruses will be used to evaluate patterns of horizontal gene transfer and evolution. Since viruses tend to be host-specific predators, it is expected that they can drive shifts community structure by selecting for resistant strains. Although this phenomenon is common in laboratory chemostats, it has not been shown definitively for natural communities. Proteomics will be used to decipher metabolic interplay through monitoring of protein production before, during, and after viral infection, while molecular methods will be used to correlate these findings with changes in the community membership. We will also examine the hypothesis that CRISPR-associated Cas proteins, which may represent a component of a microbial immune system for viral defense, are highly produced in response to virus exposure and test the prediction that the genomes of immune strains encode small RNAs complementary to phage genes. Because viral predation can undermine microbially-based technologies such as microbial fuel cells, bioethanol production, and environmental remediation, methods to study microbe-virus interactions are widely relevant to DOE missions.

Viruses play important roles in biogeochemical cycling, horizontal gene transfer, and defining the community composition of their hosts. However, we are only beginning to understand the identity and diversity of viruses in the environment. Metagenomic sequencing has recently been used to examine viral communities from a variety of environments, including seawater, marine sediment, soil, and human feces. These studies have revealed a high degree of novelty and diversity amongst environmental viruses.

The extremely high diversity of viral communities has impeded studies of virus-host interactions in natural systems. The low microbial diversity of the acid mine drainage (AMD) community from Iron Mountain, CA provides an ideal setting for studying virus-host interactions. The bacterial and archaeal communities at the Iron Mountain AMD site have already been extensively characterized, leading to the identification of virus-derived spacer sequences in the CRISPR loci (see Sun et al. poster). An important next step is to examine purified virus particles from the AMD biofilm in order to compare CRISPR loci to coexisting viruses.

Here we describe methods for viral purification from an AMD biofilm, and an initial characterization of these viruses. A combination of filtration and density-dependent centrifugation was successfully used to purify virus particles from the AMD biofilm. Transmission electron microscopy of these viruses revealed icosahedral capids typical of bacteriophage, as well as unusual morphologies similar to archaeal viruses from hot springs and other extreme environments. In addition, 3-D reconstructions using cryo-electron tomography of lower abundance ultra-small archaea (ARMAN) from AMD biofilms revealed several cells under phage attack. All of the infected ARMAN cells have altered ultrastructure, not seen in uninfected cells, including very electron dense cytoplasmic proteins. Metagenomic sequencing of the viral nucleic acids is currently underway. Detailed analysis of the viruses from the AMD biofilm will allow us to determine if viral diversity correlates with host diversity, test "Kill-the-Winner" dynamics in a natural environment, further describe the role of CRISPRs in viral resistance, understand temporal and spatial variation of virus-host systems, and gain a deeper appreciation for phage and archaeal virus diversity and evolution.

This research was supported by a grant from the DOE Genomics:GTL program (2007).

* Presenting author

109

GTL

Development of Mass Spectrometry and Proteome Informatic Approaches to Analyze the Role of Virus–Microbe Interactions in Natural Microbial Communities

Nathan VerBerkmoes[1]* (verberkmoesn@ornl.gov), Alison Russell,[1] Chongle Pan,[1] Robert L. Hettich,[1] Manesh Shah,[1] Kim Pause,[2] Mya Breitbart,[2] Christine Sun,[3] Brian Thomas,[3] Paul Wilmes,[3] Anders Andersson,[3] and Jillian Banfield[3]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [2]University of South Florida, St. Petersburg, Florida; and [3]University of California, Berkeley, California

Project Goals: See goals for abstract 109.

Microorganisms comprise the majority of extant life forms and play key roles in a wide variety of health and environmental processes, yet little is known about the nature and driving forces of their diversification. Although the roles of viruses in microbial evolution are widely recognized, neither the details of viral-microbial interactions nor the impact of virus on microbial community structure are well understood. Community genomic and proteomic (proteogenomic) methods have been established for analyzing the roles and activities of uncultivated bacteria and archaea in natural multi-species consortia (Ram, *Science* 2005, Lo *Nature* 2007, Markert, *Science*, 2007). Notably lacking are methods for monitoring viral activity in communities, tracking virus predation, and determining the consequences of viral predation for ecosystem structure. This is broadly important because it is well established that viruses can control microbial abundances, influence microbial evolution, contribute to microbial adaptation by transferring metabolic genes, and have profound effects on carbon, nutrient, and metal cycling. Studies are mostly limited by the complexity of natural consortia, which makes it difficult to observe and correlate changes in virus and host community structure over time. Our work focuses on acid mine drainage (AMD) biofilms because of their relative simplicity and established utility as a model system for development of methods for cultivation-independent analyses. Our goal is to develop proteomics and informatic techniques to characterize the microbial response to viral attacks as well as to identify the viral proteins.

Our general method for proteome characterization of microbial communities has been well established and tested. It was originally developed on the model AMD system but has been extended to other microbial communities including waste water sludge samples, enriched microbial communities from oceans and ground waters, human gut microflora, and other complex systems. The proteomics pipeline involves non-biased cell lyses via chemical methods and/or sonication, protein denaturation and reduction, protein digestion via sequencing grade trypsin. The complex peptide mixture is then separated via two dimensional chromatography (SCX-RP) coupled on-line with rapid scanning electrospray mass spectrometers (LTQs and LTQ-Orbitraps). The peptide masses and MS/MS spectra (sequencing information) are compared to protein sequencing predicted from genomic sequences via search algorithms such as SEQUEST. Peptides are identified and then matched with protein(s) they originated from.

While these methods have been shown to work well with microbial communities, even if only moderately closely related reference genomes exist, there may be new challenges for identification of viral proteins. Genomic analyses that reconstructed population genomic datasets for five complex AMD virus populations that predate bacteria and archaea (as well as for a large plasmid/phage and for many partially sampled viral populations) revealed extraordinary levels of sequence diversity, as well as rapid shuffling of sequence motifs by recombination events. This will complicate proteomic analyses because enough peptides may differ from the sequenced peptides that protein identification will be precluded. In addition, many viral proteins may be present at abundance levels that are so low that they will be difficult to identify by standard methods.

Our primary objective is to develop new proteomic approaches to integrate analyses of virus-microorganism interactions into studies of the structure and dynamics of DOE-relevant microbial communities. The main challenge for using proteomics to study viruses and microbial-viral interactions is likely to be peptide (and thus protein) identification. As noted recently, database-searching programs (such as SEQUEST) identify peptides by "looking up the answer in the back of the book" (Sadygov et al. 2004; where the genome sequence is the answer section of the book). This approach has worked extremely well for genomically-characterized microbial isolates. It also works adequately for organisms whose sequences are close to the genomically characterized type because proteins can be identified using the subset of peptides that do not contain amino acid substitutions (Denef, *JPR* 2007). In cases where multiple candidate

* Presenting author

peptide sequences are available, we can distinguish the peptides and thus identify the protein variants (Lo et al., 2007). The largest challenge for proteomics of natural viral consortia is likely to arise when candidate virus peptide sequences are not available (i.e., the virus differs in sequence from any genomically characterized virus). Little is known about population level heterogeneity in viruses in the environment. Given their fast evolutionary rates, it is likely that pure database searching methods will be inadequate. As genomic re-sequencing of every sample is not practical, we will augment database searching with de novo sequencing (sequence tagging) methods.

The first method that we have adopted to deal with the challenge of virus sequence heterogeneity is to predict proteins in all reading frames from sequence reads that are assembled into composite virus genome fragments. This step will generate a much more extensive database of reference sequences for peptide and protein identification, with the added advantage that identified peptides will assist in identification of the correct ORF calls. Another component involves modification of protocols for enzymatic protein digestion so as to produce smaller peptides that will be less likely to differ from predicted protein sequences. Finally, and most importantly, we are investigating *de novo* sequencing programs that aim to derive complete or partial amino acid sequences from MS/MS scans without complete information from protein sequence databases. Although *de novo* sequencing approaches seem straightforward, they have not been widely applied in proteomics due to low data quality and software limitations. We have optimized methods for collection of high-resolution MS/MS scans with the LTQ-Orbitrap and are currently testing the existing *de novo* programs, including PEAKS and PepNovo (Frank, *Anal Chem* 2005) for their ability to correctly *de novo* sequence known peptides. Due to limitations of each of these programs we have designed an in house *de-novo* sequencing program specifically designed to use high resolution Orbitrap MS/MS data. Initial results indicate that the program has high accuracy. We are currently testing this program using microbial community proteome data from samples known to contain a bacterium with one of the two available genome sequences. To address the challenge of viral protein abundance, we are utilizing the well studied and understood *E. coli* / MS2 phage system. The goal of this study is to determine the effect of infection on the *E. coli* proteome and to determine the ability of proteomics to detect known viral protein sequences during the course of infection. In addition, we will characterize the effects of the MS2 phage on the *E. coli* proteome over a time series from initial infection to final death of the *E. coli* culture. Once we have resolved challenges associated with sequence variation and protein abundance levels, proteomics will be used to characterize natural biofilm communities, laboratory bioreactor communities, and laboratory co-cultures and microbial isolates. We will characterize these systems before, during and after viral infection and conduct time series experiments to monitor virus and microbial interactions and co-evolution.

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

110

GTL

Community Proteogenomic Studies of Virus–Microbe Interactions in Natural Systems

Christine Sun[1]* (christine_sun@berkeley.edu), Anders Andersson,[1] Paul Wilmes,[1] Mya Breitbart,[2] Chongle Pan,[3] Brian Thomas,[1] Nathan VerBerkmoes,[3] Robert Hettich,[3] and **Jillian F. Banfield**[1]

[1]University of California, Berkeley, California; [2]University of South Florida, Saint Petersburg, Florida; and [3]Oak Ridge National Laboratory, Oak Ridge, Tennessee

Project Goals: See goals for abstract 109.

Viruses (archaeal viruses and bacteriophages) play central roles in microbial evolution and, through predation of their microbial hosts, shape the composition and functionality of ecosystems. Yet, little knowledge exists about the extent of viral population diversity and the dynamics of virus-host interactions in natural communities. The recent discovery that genomic clustered regularly interspaced short palindromic repeat (CRISPR) loci of bacteria and archaea encode virus-derived spacer sequences that provide acquired viral resistance presents a unique opportunity to examine virus-host interactions. While most natural environments harbor diverse microbial and viral populations, a model natural microbial community—the low diversity, acid mine drainage (AMD) community from Iron Mountain, CA—provides the opportunity to comprehensively examine the dynamics between CRISPR regions and viruses.

CRISPR loci within the genomes of bacteria and archaea assembled from community genomic datasets for AMD biofilms typically demonstrate high levels of heterogeneity in terms of their spacer sequence complements, resulting from spacer gain and loss. Spacers extracted *in*

* Presenting author

silico from the CRISPR loci were used to identify viral sequences. Viral genomes assembled from spacer-bearing sequences exhibit evidence of very extensive recombination, with a level of sequence rearrangement sufficient to enable viruses to elude targeting by CRISPR-encoded spacer sequences. Viruses can be linked with their hosts through the CRISPR loci, making it possible to determine host ranges. Because CRISPRs incorporate new spacers with sequences that match coexisting viruses in a unidirectional manner, the loci provide a record of recent viral exposure.

A limitation for studies of the role of viruses in the ecology of natural communities is the under-sampling of certain virus types by the CRISPR loci and the limited number of CRISPR spacers that can be recovered for each organism. To address the first challenge, we have generated virus concentrates from natural samples and are currently characterizing these by electron microscopy and genomics (see USF poster). Secondly, to expand upon the set of CRISPR spacers available for identification of viral sequences in community genomic datasets, an extensive spacer database was created for the dominant AMD bacterial organism (*Leptospirillum* group II) via 454 FLX pyrosequencing. A large number of these spacers map to a reconstructed 56 kb genome of a virus (AMDV1) inferred to target *Leptospirillum* group II. Despite over 500,000 total spacers sequences (~17,000 of them unique), rarefaction curves did not saturate, indicating rapid and extensive divergence of CRISPR loci within natural populations.

The reconstruction of sequences from natural populations of viruses provides the opportunity to generate databases of encoded proteins. However, the challenges for proteomics are considerable. At this early stage in the project, we have focused on the development of methods for protein identification that address the challenge of virus-to-virus sequence divergence. Most promising from the proteomics standpoint are *de novo* peptide sequencing methods (see ORNL poster). In ongoing work, we will evaluate the ways in which the CRISPR machinery target and sample genomes of their viral predators, examine questions related to spatial and temporal variation in viral and microbial communities, and probe the physiological interactions that occur between bacteria, archaea, and virus populations, both in laboratory biofilm cultures and in the natural environment.

This research was supported by a grant from the DOE Genomics:GTL program (2007).

111

GTL

Molecular Signatures of the Past

Elijah Roberts[1]* (erobert3@scs.uicu.edu), Jonathan Montoya,[2] Anurag Sethi,[2] **Carl R. Woese**,[3] and Zaida Luthey-Schulten[1,2,3]

[1]Center for Biophysics and Computational Biology, [2]Department of Chemistry, and [3]Institute for Genomic Biology, University of Illinois, Urbana, Illinois

Project Goals: Ribosomal signatures, idiosyncrasies in the ribosomal RNA (rRNA) and/or proteins, are characteristic of the individual domains of life. From these studies, we propose the ribosomal signatures are remnants of an evolutionary phase transition that occurred as the cell lineages began to coalesce and should be correlated with signatures throughout the fabric of the cell and its genome.

Ribosomal signatures, idiosyncrasies in the ribosomal RNA (rRNA) and/or proteins, are characteristic of the individual domains of life. As such, insight into the evolution of the modern cell can be gained from a multidimensional comparative analysis of their manifestation in the translational apparatus. In this work, we identify signatures in both the sequence and structure of the rRNA, analyze their contributions to the signal of the universal phylogenetic tree using both sequence- and structure based methods, and correlate these RNA signatures to differences in the ribosomal proteins between the domains of life. Domain specific ribosomal proteins can be considered signatures in their own right and we present evidence that they evolved at the same time as the signatures in the ribosomal RNA and therefore should not be considered recent inventions. Furthermore, we demonstrate that signatures in the rRNA coevolved with the universal ribosomal protein S4. Given S4's role in the decoding center of the ribosome, this coevolution suggests a method by which evolution may use the ribosomal proteins to fine-tune translation in different environments. From these studies, we propose the ribosomal signatures are remnants of an evolutionary phase transition that occurred as the cell lineages began to coalesce and should be correlated with signatures throughout the fabric of the cell and its genome.

* Presenting author

112

GTL

Gene Synthesis by Circular Assembly Amplification

Duhee Bang* (dbang@genetics.med.harvard.edu) and **George M. Church**

Department of Genetics, Harvard Medical School, Boston, Massachusetts

We developed a novel gene-synthesis technology[1] to effectively reduce gene synthesis error rates by a factor of ten compared to conventional methods. Gene synthesis is playing an increasingly important role in biological research. Commonly employed methods however are highly prone to errors due to the errors originating in the synthetic oligonucleotides. In our approach, exonuclease-resistant circular DNA is first constructed via the simultaneous ligation of oligonucleotides. Subsequent exonuclease degradation of the resulting ligation mixture eliminates error-rich linear products, thereby significantly improving gene-synthesis quality. By combining circular assembly amplification with the use of mismatch cleaving endonucleases we have achieved error rates of 0.025%. The method has been used to construct genes encoding Dpo4, a small thermo-stable DNA polymerase, and to construct highly repetitive DNA sequences which are not amenable to traditional synthesis methods. By adapting a uridine-cleavage strategy, we have also used the circular assembly amplification to synthesize large (>4 kb) constructs. This method promises to significantly cut the cost of gene synthesis, as the assembly of ~1kb gene (an average length of a gene) can be done in single cycle with a smaller amount of sequencing required to find a perfect construct.

Reference

1. Bang D, Church GM. Gene synthesis by circular assembly amplification. Nat Methods. 2007 Nov 25 (published online).

113

GTL

Mycoplasma Genome Synthesis and Transplantation: Progress on Constructing a Synthetic Cell

Daniel G. Gibson* (dgibson@jcvi.org), Carole Lartigue, Gwynedd A. Benders, John I. Glass, Clyde A. Hutchison III, Hamilton O. Smith, and **J. Craig Venter**

The J. Craig Venter Institute, Rockville, Maryland

Project Goals: Synthetic cell production.

Mycoplasma genitalium is an approximately 300nm diameter wall-less bacterium that has the smallest known genome of any cell that can be grown in pure culture in the laboratory. When grown under ideal conditions in a rich, serum-containing medium, as many as 100 genes appear to be dispensable based on one-gene-at-a-time transposon mutagenesis. In order to better understand the essence of a minimal cell, we are employing a synthetic genomics approach to construct a 582,970 bp *M. genitalium* genome. The synthetic genome will contain all the genes of wild type *M. genitalium* G37 except MG408, which will be disrupted by an antibiotic resistance marker to block pathogenicity and to allow for selection. Overlapping "cassettes" of 5-7 kb, assembled from chemically synthesized oligonucleotides, are being joined by *in vitro* recombination to produce intermediate assemblies of approximately 24 kb, 72 kb ("1/8 genome"), and 144 kb ("1/4 genome") and cloned as bacterial artificial chromosomes (BACs) in *Escherichia coli*. Once assemblies of all four 1/4 genomes are identified, the complete synthetic genome will be assembled and cloned in the yeast *Saccharomyces cerevisiae*. Minimization of the synthetic genome can be carried out by assembly of cassettes with individual genes deleted or by genome reduction with recombineering methods. Both approaches require the development of methods to transplant the synthesized genome into a receptive cytoplasm such that the donor genome becomes installed as the new operating system of the cell. As a step toward propagation of synthetic genomes, we completely replaced the genome of *M. capricolum* with one from *M. mycoides* LC by transplanting a whole genome as naked DNA. These cells that result from genome transplantation are phenotypically identical to the *M. mycoides* LC donor strain as judged by several criteria. The methods described are fundamental to the full development of synthetic biology.

* Presenting author

Molecular Interactions and Protein Complexes

114

GTL

The MAGGIE Project: Identification and Characterization of Native Protein Complexes and Modified Proteins from *Pyrococcus furiosus*

Angeli Lal Menon[1]* (almenon@uga.edu), Aleks Cvetkovic,[1] Sarat Shanmukh,[1] Farris L. Poole II,[1] Joseph Scott,[1] Ewa Kalisiak,[2] Sunia Trauger,[2] Gary Siuzdak,[2] and **Michael W.W. Adams**[1]

[1]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, Georgia; [2]Center for Mass Spectrometry and Department of Molecular Biology, The Scripps Research Institute, La Jolla, California

Project Goals: 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) analyses to identify PCs and MPs in the native biomass of *P. furiosus*. 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 structurally characterized. This research is part of the MAGGIE project, the overall goal of which is to determine how Molecular Assemblies, Genes, and Genomics are Integrated Efficiently. The MAGGIE project is funded by the Genomics:GTL program of DOE with John Tainer, Scripps/LBL, as the PI.

http://masspec.scripps.edu/maggie/bacteria.php

Most biological processes are carried out by dynamic molecular assemblies or protein complexes (PCs), many of which include modified proteins (MPs) containing organic and/or inorganic cofactors. The composition and the protein components of these complexes are largely unknown. They cannot be predicted solely from bioinformatics analyses, nor are there well defined techniques currently available to unequivocally identify PCs or MPs. Directly determining the identity of PCs and MPs in native biomass can resolve some of these issues. We are currently using *Pyrococcus furiosus*, a hyperthermophilic archaeon that grows optimally at 100°C, as the model organism. Fractionation of native biomass close to 80°C below the optimal growth temperature using non-denaturing, chromatography techniques, should enable purification of both stable and dynamic PCs and MPs for further characterization.

Large scale fractionation of native *P. furiosus* biomass was carried out under anaerobic and reducing conditions. Cytoplasmic proteins were fractionated by ion-exchange chromatography generating 126 fractions. These were combined into fifteen pools and subjected to 25 additional non-denaturing, chromatographic columns generating a total of 1276 fractions from all 26 chromatography steps. The identification of potential native PCs and MPs was accomplished by analyzing fractions using native and denaturing PAGE, mass spectrometry (MS), bioinformatics and for protein and metal concentrations. To date, 967 distinct *P. furiosus* proteins have been identified from the cytoplasmic protein fractionation. Based on the co-elution of proteins encoded by adjacent genes (same DNA strand), 243 proteins are proposed to be contained in 106 potential heteromeric PCs. In addition, proteins in highly-salt washed membranes were detergent-solubilzed and further fractionated using ion-exchange chromatography. A total of 205 proteins were identified from bands on PAGE. Half of them contained potential transmembrane domains, and 77 are proposed to form 25 potential PCs, which include a complex of hydrogenase (12 subunits) and ATP synthase (9 subunits).

Chromotographic fractions were also analyzed by ICP-MS for 54 elements. A subset of the elements that were unambiguously identified in the first column fractions were selected for preliminary ICP-MS analysis of the fractions from the 15 second level chromatography columns. As examples, 12 Fe-, 12 Co- and 5 Zn-containing peaks from the first ion-exchange column were resolved into 48 Fe-, 39 Co- and 13 Zn-containing peaks, respectively, on the second level columns. Using criteria such as coincident protein LC/MS/MS profiles, the presence of putative metal or metal-cofactor binding motifs and homology to characterized metalloproteins, several metal peaks in these element profiles have been assigned to known metalloproteins, while others have been identified as being potential metalloproteins. As examples, 2 Fe- and 2 Co-containing known metalloproteins were identified in the first column fractions, while at least 2 Fe-, 7 Co- and 4 Zn-containing peaks were assigned to as yet uncharacterized proteins, which are potentially

* Presenting author

novel metalloproteins. In addition, a large number of metal peaks were observed that could not be assigned to any of the proteins identified by LC/MS/MS analyses.

The results of further characterization of the soluble and membrane-bound PCs identified in this study, and particularly those containing metals, will be discussed.

115

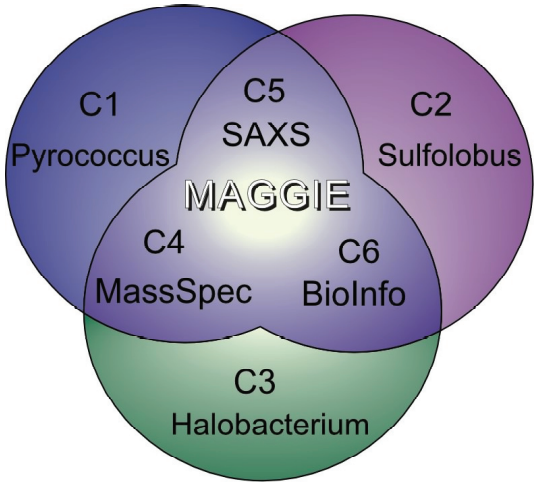
GTL

Molecular Assemblies, Genes, and Genomics Integrated Efficiently: MAGGIE

John A. Tainer[1,3]* (jat@scripps.edu), Greg L. Hura[1], Steven M. Yannone[1], Stephen R. Holbrook[1], Jane Tanamachi[1], Mike Adams[2], Gary Siuzdak,[3] and Nitin S. Baliga[4]

[1]Life Science and Physical Biosciences Divisions, Lawrence Berkeley National Laboratory, Berkeley, California; [2]University of Georgia, Athens, Georgia; [3]The Scripps Research Institute, La Jolla, California; and [4]Institute for Systems Biology, Seattle, Washington

Project Goals: MAGGIE (Molecular Assemblies, Genes, and Genomics Integrated Efficiently) will provide robust GTL technologies and comprehensive characterizations to efficiently couple gene sequences and genomic analyses with protein interactions and thereby elucidate functional relationships and pathways. To accomplish its goals, MAGGIE integrates an interdisciplinary team at Lawrence Berkeley National Lab with researchers at The Scripps Research Institute, the University of Georgia, the University of California Berkeley, and the Institute for Systems Biology into a unified Genomics:GTL program. MAGGIE thus focuses on providing an integrated, multi-disciplinary program and synchrotron facilities at the Advanced Light Source (ALS) to achieve efficient key technologies and databases for the molecular-level understanding of the dynamic macromolecular machines that underlie all of microbial cell biology. Three overall goals are 1) to facilitate instrument and technology development and optimizations though cross-disciplinary collaborations, 2) to comprehensively characterize complex molecular machines including protein complexes (PCs) and modified proteins (MPs) and 3) to provide critical enabling technologies and a prototypical map of PCs and MPs for the GTL Program. In concert, MAGGIE investigators will help to characterize microbial metabolic modularity and to provide the informed basis to design functional islands suitable to transform microbes for specific DOE missions.



MAGGIE integrates teams at Lawrence Berkeley National Lab and the Advanced Light Source (ALS) with researchers at the Scripps Research Institute, the University of Georgia, the University of California Berkeley, and the Institute for Systems Biology to achieve a molecular-level understanding of the dynamic macromolecular machines that underlie all of microbial cell biology. MAGGIE is providing improved 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 objectives can be succinctly stated: protein functional relationships can be characterized as interaction mosaics that self-assemble from independent protein pieces and that are tuned by modifications and metabolites.

MAGGIE builds strong synergies among the program components to address long term and immediate GTL objectives by combining the advantages of specific microbial systems with those of advanced technologies. A key objective for the MAGGIE Program is therefore to characterize the Protein Complexes (PCs) and Modified Proteins (MPs) underlying microbial cell biology including responses to environment. A compelling overall goal is to help reduce the 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.

MAGGIE investigators are working in concert to address GTL missions by accomplishing three specific goals: 1) develop and apply advanced mass spectroscopy and Small Angle X-ray Scattering (SAXS) technologies

* Presenting author

for high throughput characterizations of complex molecular machines including PCs and MPs, 2) create and test efficient mathematical and computational systems biology descriptions for protein functional interactions, and 3) provide both critical enabling technologies and a prototypical map of PCs and MPs for the GTL Program. Emerging databases are linked via the open-source Gaggle software system, which provides the efficient, flexible technology for communications across databases: http://gaggle.systemsbiology.org/docs/. Our new SAXS methodological treatise characterizes data interpretation tools to examine molecular interactions, flexibility, and conformational changes in solution relevant to understanding and predictions: http://bl1231.als.lbl.gov/2007/10/review_of_saxs_combined_with_c.php. The MAGGIE results are generally accessible on our website: http://masspec.scripps.edu/MAGGIE/index.php.

116

GTL

The MAGGIE Project: Production and Isolation of Tagged Native/ Recombinant Multiprotein Complexes and Modified Proteins from Hyperthermophilic *Sulfolobus solfataricus*

Stephanie Patterson,[1] Jill Fuss,[1] Kenneth Stedman,[2] Michael W.W. Adams,[3] Gary Siuzdak,[4] Trent Northen,[4] Ewa Kalisiak,[4] Sunia Trauger,[4] Nitin S. Baliga,[5] Stephen R. Holbrook,[1] John A. Tainer,[1,6] and **Steven M. Yannone**[1]* (SMYannone@lbl.gov)

[1]Department of Molecular Biology, Lawrence Berkeley National Laboratory, Berkeley, California; [2]Center for Life in Extreme Environments, Portland State University, Portland, Oregon; [3]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, Georgia; [4]Center for Mass Spectrometry, The Scripps Research Institute, La Jolla, California; [5]Institute for Systems Biology, Seattle, Washington; and [6]Department of Biochemistry and Molecular Biology, The Scripps Research Institute, La Jolla, California

Project Goals: As part of MAGGIE, we are developing high throughput recombinant DNA and native biomass technologies for the extremophilic organism *Sulfolobus solfataricus* which has a growth optimum at 80° C and pH 3.0. We are also exploiting the unique characteristics of Archaeal membranes to isolate membrane-protein as well as soluble protein complexes from native biomass. We have developed universally applicable approaches for enriching native biomass for protein complexes and established simplified proteomic analyses resulting in greater than 50% coverage of the *Sulfolobus* proteome with relatively simple and rapid procedures. We are developing the computational tools necessary to integrate protein fractionation, predicted/observed molecular mass, genomic organization, and co-fractionation data sets to identify novel protein complexes. Using MS/MS-based metabolomic analyses, we have identified an un-annotated trehalose synthesis pathway in *Sulfolobus* and are expanding this approach to validate homology-based functional predictions. We have begun the structural characterization of recombinant *Sulfolobus* proteins with small angle x-ray scattering (SAXS) at the advanced light source at LBNL, and will discuss the biological implications of these studies. We are testing the idea that the hyperthermophilic nature of *Sulfolobus* will allow us to "thermally trap" protein complexes assembled at 80° C by isolating these complexes at room temperature.

Dynamic protein-protein interactions are fundamental to most biological processes and essential for maintaining homeostasis within all living organisms. Understanding the networks of these protein interactions is of critical importance to understanding the complexities of biological systems. The MAGGIE project was conceived, in part, as a response to the DOE GTL initiative to develop technologies to map the proteomes of model organisms. In this project we are exploiting unique characteristics of members of extremophilic Archaea to identify, isolate, and characterize multi-protein molecular machines. We have teamed expertise in mass spectrometry, systems biology, structural biology, biochemistry, and molecular biology to approach the challenges of mapping relatively simple proteomes.

As part of MAGGIE, we are developing high throughput recombinant DNA technologies for the extremophilic organism *Sulfolobus solfataricus* which has a growth optimum at 80°C and pH 3.0. We are using a naturally occurring viral pathogen of this organism to engineer shuttle vectors designed for recombinant protein tagging and expression in the native *Sulfolobus* background. We are also exploiting the unique characteristics of *Archaeal* membranes to isolate membrane-protein as well as soluble protein complexes from native biomass. We have developed universally applicable approaches for enriching native biomass for protein complexes and established simplified proteomic analyses resulting in greater than 50% coverage of the *Sulfolobus* proteome with relatively simple and rapid procedures. We are developing the computational tools necessary to integrate protein frac-

* Presenting author

tionation, predicted/observed molecular mass, genomic organization, and co-fractionation data sets to identify novel protein complexes. Using MS/MS-based metabolomic analyses, we have identified an un-annotated trehalose synthesis pathway in *Sulfolobus* and are expanding this approach to validate homology-based functional predictions. We have begun the structural characterization of recombinant *Sulfolobus* proteins with small angle x-ray scattering (SAXS) at the advanced light source at LBNL, and will discuss the biological implications of these studies. We are testing the idea that the hyperthermophilic nature of *Sulfolobus* will allow us to "thermally trap" protein complexes assembled at 80°C by isolating these complexes at room temperature. Ultimately, we aim to identify metabolic modules suitable to transfer specific metabolic processes between microbes to address specific DOE missions while developing generally applicable molecular and biophysical technologies for GTL.

117

GTL

Metabolomic Profiling of a Hyperthermophile and the Characterization of Metabolite-Protein Interactions

Sunia A. Trauger[1]* (strauger@scripps.edu), Ewa Kalisak,[1] Jaroslaw Kalisiak,[1] Hirotoshi Morit,[1] Michael V. Weinberg,[2] Angeli Lal Menon,[2] Farris L. Poole II,[2] Michael W.W. Adams,[2] and Gary Siuzdak[1]

[1]Scripps Center for Mass Spectrometry and the Departments of Molecular Biology and Chemistry, The Scripps Research Institute, La Jolla, California and [2]The Department of Biochemistry and Molecular Biology, University of Georgia, Athens, Georgia

We have performed a comprehensive characterization of global molecular changes using the hyperthermophilic archaeon, *Pyrococcus furiosus,* as a model organism and using transcriptomic (DNA microarray), proteomic and metabolomic analysis as it undergoes a cold adaptation response from its optimal 95°C to 72°C. Metabolic profiling on the same set of samples show the down-regulation of many metabolites. However, some metabolites are found to be strongly up-regulated. An approach using accurate mass, isotopic pattern, database searching and retention time is used to putatively identify several metabolites of interest. Many of the up-regulated metabolites are part of an alternative polyamine biosynthesis pathway previously established in a thermophilic bacterium *Thermus thermophilus*. [1]

Arginine, agmatine, spermidine and branched polyamines N^4-aminopropylspermidine and N^4-(*N*-acetylaminopropyl)spermidine were unambiguously identified based on their accurate mass, isotopic pattern and matching of MS/MS data acquired under identical conditions for the natural metabolite and a high purity standard. For the branched polyamines N^4-aminopropylspermidine and N^4-(*N*-acetylaminopropyl)spermidine, both DNA microarray and semi-quantitative proteomic analysis using a label-free spectral counting approach indicate the down-regulation of a large majority of genes with diverse predicted functions related to growth such as transcription, amino acid biosynthesis and translation. Some genes are however, found to be up-regulated through the measurement of their relative mRNA and protein levels. A novel approach using metabolite immobilization for protein capture followed by proteomic analysis is used for the identification of protein partners which may interact with three polyamines. These were agmatine, spermidine and the novel metabolite N^4-(*N*-acetylaminopropylsprmidine) involved in the alternative polyamine biosynthetic pathway. Proteins identified using this method included unique proteins, as well as ones which were common to all three polyamines. Proteins identified using immobilized spermidine as bait, included SAM decarboxylase SpeD (PF1930) and S-adenosylmethionine synthetase (PF1866) which are directly involved in its probable biosynthetic pathway. Other proteins identified with spermidine immobilization are involved in translation such as PF1375 (translation elongation faction ef1), PF1367 (LSU ribosomal protein L7AE), PF1264 (translation elongation factor eIF-5a). Polyamines such as spermidine are known to play a critical role in hyperthermophiles in translation, ribosomal assembly and protein elongation and their interaction may reflect the specific affinity of spermidine for these proteins. For the novel metabolite N^4-(*N*-acetylaminopropylsprmidine), for which the enzymatic pathways involved in its synthesis remain unknown, a conserved hypothetical protein PF0607 was uniquely identified. While, proteins identified include those which are clearly interacting with other proteins and DNA that bind the immobilized metabolite, we believe this is a promising technique which could be used as an initial screen for uncovering enzymatic processes that underlie the biosynthesis of newly identified metabolites. The complimentary information obtained by the various 'omics' techniques are used to catalogue and correlate the overall molecular changes.

Reference

1. Ohnuma, M.; Terui, Y.; Tamakoshi, M.; Mitome, H.; Niitsu, M.; Samejima, K.; Kawashima, E.; Oshima, T., N1-aminopropylagmatine, a new

* Presenting author

polyamine produced as a key intermediate in polyamine biosynthesis of an extreme thermophile, Thermus thermophilus. *J Biol Chem* **2005,** 280, (34), 30073-82.

118

GTL

Protein Complex Analysis Project (PCAP): Project Overview

Lauren Camp,[1] Swapnil Chhabra,[1] Dwayne Elias,[3] Jil T. Geller,[1] Hoi-Ying Holman,[1] Dominique Joyner,[1] Jay Keasling,[1,2] Aindrila Mukhopadhyay,[1] Mary Singer,[1] Tamas Torok,[1] Judy Wall,[3] **Terry C. Hazen**,[1*] Simon Allen,[4] Gareth Butland,[1] Megan Choi,[1] Ming Dong,[1] Barbara Gold,[1] Steven C. Hall,[4] Bing K. Jap,[1] Jian Jin,[1] Susan J. Fisher,[4] Haichuan Liu,[4] Ramadevi Prathapam,[1] Evelin Szakal,[4] Peter J. Walian,[1] H. Ewa Witkowska,[4] Lee Yang,[1] Wenhong Yang,[1] **Mark D. Biggin**[1*] (mdbiggin@lbl.gov), Pablo Arbelaez,[1] Manfred Auer,[1] David Ball,[1] Myhanh Duong,[1] Robert M. Glaeser,[1] Bong-Gyoon Han,[1] Danielle Jorgens,[1] Jitendra Malik,[2] Hildur Palsdottir,[1] Jonathan P. Remis,[1] Dieter Typke,[1] **Kenneth H. Downing**,[1*] Steven S. Andrews,[1] Adam P. Arkin,[1,2] Steven E. Brenner,[1,2] Y. Wayne Huang,[1] Keith Keller,[1] Ralph Santos,[1] Max Shatsky,[1,2] and **John-Marc Chandonia**[1*]

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

Project Goals: 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 another Genomics:GTL initiative conducted by the Environmental Stress Pathways Project (ESPP). ESPP has developed *D. vulgaris* as a model for stress responses and has 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 are extending ESPP's 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 work of the ESPP 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 is the development of automated high throughput methods that are robust and allow for the comprehensive analysis of many protein complexes. Biochemical purification of endogenous complexes and identification by mass spectrometry is being coupled with in vitro and in vivo EM molecular imaging methods. Because no single method can isolate all complexes, we are developing two protein purification pipelines, one the current standard Tandem Affinity Purification approach, the other a novel tagless strategy. Specific variants of each of these are being developed for water soluble and membrane proteins. Our Bioinstrumentation group is developing highly parallel micro-scale protein purification and protein sample preparation platforms, and mass spectrometry data analysis is being automated to allow the throughput required. The stoichiometries of the purified complexes are being determined and the quaternary structures of complexes larger than 250 kDa are being solved by single particle EM. We are developing EM tomography approaches to examine whole cells and sectioned, stained material to detect complexes in cells and determine their localization and structures. New image analysis methods will be applied to speed

* Presenting author

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.

Our progress during the second year of the project includes establishing both Gateway and Recombineering based pipelines for constructing genetically altered *D. vulgaris* strains; ramping up biomass production; establishing an optimized four-step tagless fractionation series for the purification of water soluble protein complexes from 400L of culture; establishing proof of principle data for the effectiveness of the tagless identification of protein complexes by mass spectrometry; scaling up tagless purification of inner and outer membrane complexes; identifying over 50 water soluble and membrane complexes; automating many aspects of mass spectrometry data analysis; establishing a TAP pipeline; determining the structure of five additional complexes by single particle EM; developing an improved automated particle picking method for EM images of purified complexes; establishing fluorescent SNAP tag labeling of complexes in biofilms; discovering differences in activity between cells associated with fibers in biofilms; and establishing a novel approach (WIST) for automated construction of database web interfaces that speeds database construction and using it to build LIMS modules to store data from several parts of our workflow. Further details on these and other results are provided in the Subproject specific posters.

119

GTL

Protein Complex Analysis Project (PCAP): Multi-Protein Complex Purification and Identification by Mass Spectrometry

Simon Allen,[2] Gareth Butland,[1] Megan Choi,[1] Ming Dong,[1] Steven C. Hall,[2] Bing K. Jap,[1] Jian Jin,[1] Susan J. Fisher,[2] Haichuan Liu,[2] Evelin D. Szakal,[2] Peter J. Walian,[1] H. Ewa Witkowska,[2] Lee Yang,[1] and **Mark D. Biggin**[1]* (mdbiggin@lbl.gov)

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

Project Goals: This subproject of the Protein Complex Analysis Project (PCAP) is developing several complementary high throughput pipelines to purify 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 PCAP's effort to model 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 TOF/TOF mass spectrometry. Our second purification approach for water soluble proteins uses and extends 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 are being designed to capture membrane protein complexes. A major part of our effort is 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 kDa are being 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 PCAP's Bioinformatics Subproject with computational models of stress response pathways that are currently being established by the Environmental Stress Pathways Project (ESPP).

Our results for the second year of the project are as follows.

* Presenting author

Tagless purification of water soluble complexes. We have developed an optimized four-step fractionation scheme for the tagless purification strategy that uses protein from 400L of culture and have used it to identify and purify over 50 homomeric and heteromeric water soluble protein complexes from just 0.2% of the fraction space. We have established an efficient, highly reproducible mass spectrometry sample preparation protocol that uses 96-well PVDF multiscreen plates and is effective with the iTRAQ methodology we have adopted to quantitate the relative abundances of polypeptides in different chromatographic fractions. Methods for preparing protein samples suitable for single particle EM analysis are being refined, including the use of different crosslinking reagents to stabilize complexes on EM grids. To date, 16 complexes have been sent to the EM Subproject for structural determination. As a result, structures at various resolution have been obtained for Pyruvate Ferredoxin Oxidoreductase, GroEL, a putative protein DVU0671, PEP synthase, and 6,7-dimethyl-8-ribityllumazine synthase.

Tagless purification of membrane complexes. Over the past year we have isolated membrane protein complexes using developmental protocols featuring several chromatographic steps (ion exchange, hydroxyapatite and molecular sieve) and blue native gel electrophoresis. For this work, mild detergents have been used to sequentially solublize proteins of the inner and outer membrane. With this approach and an improved protocol for preparing mass spectrometry samples, 20 membrane protein complexes (homomeric and heteromeric) have been identified. Membrane protein complex samples have also been prepared for preliminary electron microscopy analysis and delivered to the EM group.

Tandem Affinity Purification of water soluble complexes. We have completed trials of different TAP tag combinations for protein complex purification from *D. vulgaris*. Initial tests have compared the efficiency of the Sequential Peptide Affinity (SPA) tag and the Strep-TEV-FLAG (STF) tag. We have confirmed that both tags can purify proteins synthesized in *D. vulgaris* with comparable high yield and low background binding properties. We have also completed optimization trials to determine the quantities of *D. vulgaris* biomass required for purification of protein complexes in amounts sufficient for identification by mass spectrometry. Strains bearing individually tagged genes will be generated for this high throughput purification pipeline using high-throughput cloning strategies currently being deployed to construct tagged *D. vulgaris* genes rapidly and efficiently in *E. coli*.

Automation of protein complex purification. We have developed a multi-channel, native gel electrophoresis instrument for high resolution protein separation and automated band collection. This instrument can separate samples of protein mixtures from ~20Kd to ~600Kd and elute a protein band into a 200 µl fraction, without noticeable loss of sample. The use of this free-flow electrophoresis apparatus will greatly assist our efforts to achieve high throughput and provide an additional means of obtaining specimens in amounts appropriate for EM studies.

Mass spectrometry. We have worked in parallel both to further optimize mass spectrometry sample preparation and data acquisition and to discover *DvH* complexes separated *via* a tagless protein complex identification pipeline. We have implemented use of an internal protein standard to monitor recoveries during sample preparation of iTRAQ-labeled tryptic peptides, allowing us to normalize quantitation results. We have analyzed recoveries of peptides varying in hydrophobicity, charge and size from the PVDF membranes used in our high throughput sample preparation method. We have also demonstrated that our mass spectrometry protocols are accurate for samples of lower concentration and lower total protein load, down to 2 µg. Protein complex discovery was performed on 0.2% of the fraction space of soluble protein derived from 400 L prep (76 sizing column fractions). *In toto*, 160 polypeptides were matched to 2 or more peptides and a further 70 polypeptides matched to a single peptide. At least 7 heteromeric complexes orthologous to *E. coli* complexes, 3 novel heteromeric complexes, and 39 known homomeric complexes were identified as well as tens of additional polypeptides whose size migration suggested are part of a complex. Three of the heteromeric complexes were also examined by TAP, giving similar results to those obtained by the Tagless approach. The large number of complexes detected within a very small portion of the overall analytical space indicates that the tagless strategy holds high potential for characterizing the bacterial interactome. We have also employed Synapt HD mass spectrometer (Waters) to characterize heterogeneity within dissimilatory sulfite reductase and found that discrete forms of the complex differ in subunit stoichiometry.

Integrated mass spectrometry data acquisition and automated data analysis. To handle the mass spectrometry data for the large number of protein fractions generated by our tagless strategy, an integrated data acquisition and automated processing pipeline has been developed. This integrates commercial ProteinPilot software with several home-developed processing tools. ProteinPilot fetches data directly from the Oracle database of our

* Presenting author

AB4800 mass spec. instrument and produces lists of identified proteins and their relative concentrations. The tools we have developed generate normalized elution profiles for each detected protein and allow automatic initiation and monitoring of the protein identification process once MS/MS data are available on the Oracle database. To reduce redundant MS/MS data acquisitions and improve coverage of less abundant proteins, an iterative and intelligent data acquisition has been integrated into the pipeline. We have developed a comparator that pulls data and information from various modules in the pipeline and generates inclusion/exclusion lists to direct subsequent data acquisition. To complete the system, a dynamic instrument command control module is being developed that takes inclusion/exclusion lists and predictions of peptide retention times as inputs and produces more effective MS/MS data acquisition requests. In operation it will control and monitor AB4800's data acquisition and provide on-the-fly assessments of spectrum quality and peptide id feasibility. A clustering analysis tool has also been developed that can effectively identify protein complexes. This automatic tool takes protein elution profiles as input, performs several preprocesses including peak identification, de-noising and peak-overlap characterization, and then calculates correlation coefficients of every pair of peaks and ranks and sorts results to identify components of protein complexes.

120

GTL

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

Adam P. Arkin,[1,2] Steven E. Brenner,[1,2] Max Shatsky,[1,2] Ralph Santos,[1] Wayne Huang,[1] Keith Keller,[1] and **John-Marc Chandonia**[1,2*] (jmchandonia@lbl.gov)

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

Project Goals: 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 other projects within the Virtual Institute for Microbial Stress and Survival (VIMSS), 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. In addition to storing experimental data produced by the PCAP project, we will assess the quality and consistency of the data, and compare our results to other public databases of protein complexes, pathways, and regulatory networks. We will prioritize proteins for tagging, TAP, and study by EM based on analysis of VIMSS data and other bioinformatic predictions. 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. Ultimately, we wish to analyze PCAP data in such a way as to automatically generate hypothetical models of cellular pathways, which will be validated by comparison to experimental observations.

We are developing a modular LIMS system to store data and metadata from the high-throughput experiments undertaken by the other PCAP subprojects. Each module of the LIMS corresponds to a step in the experimental pipeline. We have developed WIST (Workflow Information Storage Toolkit), a template-based toolkit to facilitate rapid LIMS development. WIST allows LIMS programmers to design multi-step workflows using modular core components, which can be added and arranged through a simple, intuitive configuration and template mechanism. WIST uses the templates to create unified, web-based interfaces for data entry, browsing, and editing. We have deployed WIST in an updated version of the tagless purification module, the tagged purification module, and as a component of our automated pipeline for sequence validation of high throughput constructs.

We have also prioritized proteins for tagging, TAP, and study by electron microscopy based on analysis of gene expression data from the VIMSS Environmental Stress Pathway Project (ESPP) and bioinformatic predictions. To date, we have identified 1217 *D. vulgaris* proteins as high-priority targets for tagging by the PCAP Microbiology Core. 265 of these proteins have already been identified as likely components of multimeric complexes by mass spectroscopy of fractions purified using the tagless pipeline. We will cross-validate our results by comparing the composition of complexes characterized by both the tagless and tagged purification pipelines. In addition, *D. vulgaris* orthologs of proteins belonging to previously

* Presenting author

characterized complexes from other organisms have been selected as tagging targets. This was done in order to study the degree to which stable inter-protein interactions are conserved between orthologs, and to establish a baseline characterization of potential complexes to compare with the same proteins under stress conditions. We also plan to study the degree to which correlated expression in microarray experiments may be used to predict stable protein interactions.

121

GTL

Protein Complex Analysis Project (PCAP): Imaging Multi-Protein Complexes by Electron Microscopy

Manfred Auer, David Ball, Myhanh Duong, Danielle Jorgens, Hildur Palsdottir, Jonathan Remis, and **Kenneth H. Downing*** (KHDowning@lbl.gov)

Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California

Project Goals: 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. A second goal is to determine the spatial organization and relative locations of large multi-protein complexes within individual, intact microbes. 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 individual, lysed cells. 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. Advanced computational methods are being developed to enhance each of these experimental goals.

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. Our goal is to determine the spatial organization and relative locations of large multi-protein complexes within individual, intact microbes, as well as in microbial communities, using cryo-EM tomography and novel tag-based labeling approaches.

It has quite recently been established that cryo-EM tomography can be used to produce clearly distinguishable images of larger multiprotein 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. Our present aim is to characterize large subcellular structures in *Desulfovibrio vulgaris* to provide a basis for understanding the morphological changes that follow various stresses.

We also employ plastic-section electron microscopy to study both planktonic and biofilm forms of microbes of interest. This approach has the advantage that it 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.

In order to take advantage of the genetic tools that allow tagging of specific proteins for localization by both light and electron microscopy, we are focusing on several fluorescent reagents that can be characterized in the light microscope and then photoconverted to electron-dense signals for electron microscopy. This is quite a new endeavor for anaerobic bacteria such as *D. vulgaris*, which produce high levels of H_2S. Our initial focus is on morphology of biofilms in which we see a number of structures that have yet to be characterized in *D. vulgaris*. We grow biofilms of *D. vulgaris* on cellulose dialysis tubing or sheets, where the biofilms cover almost the entire available surface area. Samples are high pressure frozen and freeze-substituted in order to optimize preservation of structural details. Electron microscopic analysis of biofilm sections reveals 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. Variations in the deposition patterns indicate that metal reduction activity

* Presenting author

varies between neighboring cells in biofilms. Using microwave processing of 6-day old mature *D. vulgaris* biofilm, we have confirmed the existence of such metal strings that extend dozens and possibly hundreds of microns. Interestingly, we found that intact cells are associated with these metal strings, whereas areas devoid of such metal strings only contain cell debris, suggesting that these metals strings contribute to cell survival in such stationary biofilms.

We have developed on-grid culturing methods for rapid study of such features in cell monolayers grown under various environmental conditions, and found the presence of filamentous structures of ~ 7nm in diameter that were associated with metals precipitated out of a Uranium-containing solution.

We have tested ReAsH and SNAP-labeling of several strains of *D. vulgaris* in which proteins have been tagged by members of the PCAP Microbiology group. The SNAP-labeling appears promising as judged by light microscopy. In-invo and in-vitro labeling of tagged proteins, before and after cell lysis, respectively, followed by SDS PAGE suggests specific binding for the SNAP-tag reagent. We found large variations in the labeling intensity of planktonic cells, with only about 20-30% of these genetically identical bacteria displaying strong labeling. We have ruled out the possibility that this difference is due to variability of reagent access or vitality of the cells, suggesting that there are large differences in protein expression levels even at the planktonic state. We speculate that differences in protein expression levels may be the reason for cell-to-cell differences of metal reduction capability as seen in planktonic cells and in biofilms. We are currently optimizing the photoconversion of the fluorescence signal both for planktonic cells and biofilms.

While the intact *D. vulgaris* cells are generally thicker than optimal for high resolution electron tomography, initial results show that various approaches to specimen preparation can produce samples in which a wealth of internal detail can be visualized. We have begun to characterize the initial morphological responses to oxidative stress, which are particularly dramatic in cells that have the stored energy resources to mount a successful metabolic response.

122

GTL

Protein Complex Analysis Project (PCAP): 3-D Reconstruction of Multi-Protein Complexes by Electron Microscopy

Bong-Gyoon Han,[1] Dieter Typke,[1] Ming Dong,[1] Pablo Arbelaez,[2] Jitendra Malik,[2] Mark D. Biggin,[1] and **Robert M. Glaeser**[1]* (rmglaeser@lbl.gov)

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

Project Goals: The broad aim of this component of PCAP is to develop high-throughput capabilities for determining the overall morphology and arrangement of subunits within large, biochemically purified multi-protein complexes of *Desulfovibrio vulgaris* Hildenborough.

Three-dimensional (3-D) reconstructions are obtained by single-particle electron microscopy (EM) at a resolution of ~2 nm for either negatively stained or unstained (cryo-EM) specimens. The goals of determining the quaternary structures of multi-protein complexes include (1) determining whether structural changes occur in some molecular machines under markedly different physiological conditions, such as those that would be encountered in the field during bioremediation, (2) providing 3-D models of their structures that can be used as templates in order to image the same multi-protein complexes within whole cells by EM tomography, and (3) using both types of information to model the biochemical networks and circuits of micro-organisms in order to better utilize them for applications in bioremediation or bioenergy.

Single-particle EM within PCAP has focused during the first two years on large soluble-protein complexes with Mr in the range 400 k to over 1000 k. These complexes have been found to differ considerably in terms of how well they hold up during EM sample preparation, and not all are stable even under the currently used conditions of cryo-EM sample preparation. Roughly half of the complexes studied have been stable enough to produce high-quality 3-D reconstructions, however, and class-average projection-images have been obtained for most of the others.

This preliminary phase of characterization has shown surprising differences in the quaternary structures of complexes isolated from *Dv*H and those that are already

* Presenting author

known for homologous proteins from other microbes. These differences occur frequently enough to make it clear that structures determined for other micro-organisms are inadequate for use as templates for modeling the biochemical networks within a given microbe of interest. By extension it is clear that the same type of EM structure determinations could be essential to characterize any changes in the multi-protein complexes that exist under different physiological conditions.

Work that is aimed towards increasing the throughput of single-particle EM currently includes the implementation of automated data collection and automated data analysis, and the engineering of new support-film technologies for EM sample preparation. The latter is driven by the need, encountered within this high-throughput project, to use technologies that do not require sample-dependent optimization and are more likely to preserve quaternary structure in a conformationally homogeneous state.

123

GTL

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[1,4]* (tchazen@lbl.gov), Hoi-Ying Holman,[1,4] Jay Keasling,[1,2,4] Aindrila Mukhopadhyay,[1,4] Swapnil Chhabra,[1,4] Jil T. Geller,[1,4] Mary Singer,[1,4] Dominique Joyner,[1,4] Lauren Camp,[1,4] Tamas Torok,[1,4] Judy Wall,[3,4] Dwayne Elias,[3,4] and **Mark D. Biggin**[1,4]

[1]Lawrence Berkeley National Laboratory, Berkeley California; [2]University of California, Berkeley, California; [3]University of Missouri, Columbia, Missouri; and [4]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

Project Goals: The Microbiology Subproject of the Protein Complex Analysis Project (PCAP) provides 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 are building on techniques and facilities established by the Environmental Stress Pathways Project (ESPP) for isolating, culturing, and characterizing *Desulfovibrio vulgaris*. The appropriate stressors for study will be identified and, using stress response pathway models from ESPP, the relevance and feasibility for high throughput protein complex analyses will be assessed. Two types of genetically engineered strain are being constructed: strains expressing affinity tagged proteins and knock out mutation strains that eliminate expression of a specific gene. High throughput phenotyping of these engineered strains will then be used to determine if any show phenotypic changes. Finally, we are producing large quantities of cells under different conditions and harvesting techniques for optimal protein complex analyses.

The Microbiology Subproject of PCAP provides 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. This project has built 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 have been identified and, using stress response pathway models from VIMSS, the relevance and feasibility for high throughput protein complex analyses is being assessed. We also produce all of the genetically engineered strains for PCAP. Three types of strains are being constructed: strains expressing affinity tagged proteins, those expressing fluorescent tags for sub-cellular localization, and knock out mutation strains that eliminate expression of a specific gene. We anticipate producing several hundred strains expressing affinity tagged proteins for complex isolation and EM labeling experiments by the other Subprojects. A much smaller number of knock-out mutation strains are being produced to determine the effect of eliminating expression of components of putative stress response protein complexes. Both types of engineered strains are being generated using a two-step procedure that first integrates and then cures much of the recombinant DNA from the endogenous chromosomal location of the target gene. We are developing 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 in operons. The Microbiology Subproject provides high throughput phenotyping of all engineered strains to determine if any show phenotypic changes. We also determine if the tagged proteins remain functional and that they do not significantly affect cell growth or behavior. The knockout mutations are tested in a comprehensive set of conditions to determine their ability to respond to stress. High throughput optimiza-

* Presenting author

tion of culturing and harvesting of wild type cells and all engineered strains are 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 are producing 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 use extreme levels of 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 – 400 L of culture, and we expect more than 4,000 liters of culture will be prepared and harvested every year. The Microbiology Subproject is optimizing 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 last year, the Microbiology Subproject produced biomass for multi-protein complex isolation and identification by mass spectrometry, and for imaging multi-protein complexes by electron microscopy. This year we have provided more than 2000 L of biomass consisting of more than 300 individual productions. Production volumes range from less than 1 L of DvH wt, and mutants, for imaging and development of high-throughput tagging and isolation methods, to 400 L of DvH wt for isolation of membrane protein complexes. We currently produce 100 L of DvH wt in five days, operating two 5 L fermenters in continuous flow mode in parallel. Extensive monitoring and assays are performed to ensure product quality and consistency, including continuous measurement of optical density and redox potential, and discrete sampling for AODCs, anionic composition (including organic acids), anaerobic and aerobic plating, total protein concentrations, PLFA and qPCR. The goal of incorporating different affinity or tandem affinity (TAP) protein tags into three genes to determine the best tag for use in the PCAP project has been completed. These included the Strep-tag® (IBA) for streptavidin-binding, the SPA-tag (a.k.a. CTF) that consists of a calmodulin binding motif, tobacco etch virus protease (TEV) and 3X FLAG affinity, as well as a combination of these that replaces the calmodulin binding with the Strep-tag® resulting in STF. The three genes were the dissimilatory sulfite reductase subunit C, pyruvate ferredoxin oxidoreductase subunit B and ATP synthase subunit C. Additionally, several other gene targets have been identified through close collaboration with the VIMSS/ESPP group at LBNL and are currently being tagged. To determine localization of a given gene product in the cell, we have utilized the tetracysteine, SNAP™ (Covalys) and 6XHis tags in cooperation with the EM group of the PCAP project. Currently the total number of genes tagged with CTF are 6, with STF 17, with strep 30, with tetracysteine 13 and with SNAP 11. Tagged genes were successfully generated in randomly cloned fragments of *D. vulgaris* DNA by recombineering techniques. These are being examined for introduction of the tagged genes into the *D. vulgaris* chromosome by two recombination events. This success paves the way for a HTP tagging procedure with an ordered plasmid library of *D. vulgaris* DNA fragments. To improve the plasmid insertion tagging currently used, we implemented the two-step TOPO-GATEWAY strategy (Invitrogen) for the production of a library of 145 entry clones in *E. coli*. We also constructed a library of custom destination vectors bearing the following tags: 6xHis, STF, SPA, SNAP, STF-6xHis and STF-SNAP. These destination vectors enable rapid addition of desired tags to the entry vector library. Consequently we have constructed a library of 140 STF/SPA tagged clones of which, 84 have been electroporated in *D. vulgaris* so far. Construction of the entry vector and tagged clone libraries involved development of automated software and hardware methods. On the software end we collaborated with Subgroup D (Computational Core) for the development of automated algorithms and a LIMS system for: 1) Primer identifications based on gene locations within operons for PCR amplifications in 96-well format, and 2) QA/QC for sequence data analysis and sample tracking. On the hardware end we developed and implemented methods for handling nucleic acids using a liquid handling system from Beckman Coulter and collaborated with Subgroup B (Hardware engineering) for the design of a custom electroporation device to enable rapid transformation of tagged constructs in *D. vulgaris*. All of the tagged strains constructed this year have been characterized using phenotypic microarrays (PM), and the *D. vulgaris* megaplasmid minus strain (MP(-)) being used in the electroporation studies last year was aggressively characterized for all differences including stress responses with the wildtype and was found to have some significant response differences. This enabled the group to redirect transformation studies using electroporation away from the MP(1) strain and towards the wildtype only.

* Presenting author

124

GTL

Protein Complex Analysis Project (PCAP): High Throughput Strategies for Tagged-Strain Generation in *Desulfovibrio vulgaris*

Swapnil Chhabra[1]* (SRChhabra@lbl.gov), Gareth Butland,[1]* Dwayne Elias,[2] **Veronica Fok**,[1] **Barbara Gold**,[1] Jian Jin,[1] Aindrila Mukhopadhyay,[1] **Ramadevi Prathapam**,[1] **Wenhong Yang**,[1] John-Marc Chandonia,[1] Judy Wall,[2] Terry Hazen,[1] and Jay Keasling[1,3]

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

Project Goals: As part of the microbiology core of the Protein Complex Analysis Project (PCAP) our goal is to develop a technological platform for creating a library of *D. vulgaris* mutant strains expressing tagged proteins at high throughput. Based on the workflow designed around the TOPO-GATEWAY strategy, we will produce a hundred constructs carrying the STF tag which will be transformed in *D. vulgaris* to create a tagged strain library. We are also exploring an alternative high throughput strategy using an ordered library of *D. vulgaris*.

In this poster we describe our efforts towards the development of a high throughput platform for generating a library of *D. vulgaris* mutant strains expressing tagged proteins. This work is part of the microbiology core of the Protein Complex Analysis Project (PCAP). We highlight our efforts towards automating the strain generation process using automated software and hardware tools such as LIMS for automated sequence alignments, liquid handling systems for processing nucleic acids and custom robotics for high throughput electroporations. For generation of tagged clones we have developed and tested two approaches. The first one involves the use of plasmid constructs carrying single target genes using the two-step TOPO-GATEWAY cloning approach (Invitrogen). The first step in the strategy involves generation of an entry vector carrying the gene of interest (GOI) via TOPO cloning. The second step involves transfer of the GOI from the entry vector to a suitable destination vector (carrying the tag of choice) through an in-vitro recombination reaction. This approach works best for genes located at terminal ends of operons and based on this approach we have constructed a library of 145 tagged clones in *E. coli*.

The second strategy involves the use of an ordered library of *D. vulgaris* modified using a lambda-red phage system. Library constructs are modified, in a strain of *E. coli* expressing the lambda-red recombination system, using linear PCR products specifically engineered to recombine into the 3' end of the gene of interest. These PCR products, when inserted into the gene of interest modify the coding sequence of the gene to encode a C-terminal fusion protein bearing the tag of choice. We have performed several rounds of trials to optimize the recombineering protocol and have created 15 constructs bearing individually tagged genes. This system is now being integrated into the high-thoughput tagged strain construction pipeline. The recombineering system will work in tandem with the TOPO-GATEWAY method, and will focus on genes currently not amenable to tagging via that system. Constructs generated via both strategies have been transformed into *D. vulgaris* via electroporation and are currently being tested.

125

GTL

The Center for Molecular and Cellular Systems: Biological Insights from Large Scale Protein-Protein Interaction Studies

Michelle V. Buchanan[1]* (buchananmv@ornl.gov), Dale A. Pelletier,[1] Gregory B. Hurst,[1] W. Hayes McDonald,[1] Denise D. Schmoyer,[1] Jennifer L. Morrell-Falvey,[1] Mitchel J. Doktycz,[1] Brian S. Hooker,[2] William R. Cannon,[2] H. Steven Wiley,[2] Nagiza F. Samatova,[3] Tatiana Karpinets,[1] Mudita Singhal,[2] Chiann-Tso Lin,[2] Ronald C. Taylor,[2] Don S. Daly,[2] Kevin K. Anderson,[2] and Jason E. McDermott[2]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [2]Pacific Northwest National Laboratory, Richland, Washington; and [3]North Carolina State University, Raleigh, North Carolina

Project Goals (Abstracts 126-130): The Center for Molecular and Cellular Systems (CMCS) has established a resource for high-throughput determination of protein-protein interactions (PPI) for the Genomics:GTL community.

The Center for Molecular and Cellular Systems (CMCS) has established a resource for high-throughput determination of protein-protein interactions (PPI) for

* Presenting author

the Genomics:GTL community. As part of the CMCS, an analysis "pipeline" has been established for identifying PPI among soluble proteins in *Rhodopseudomonas palustris*. The general strategy is to express an affinity-tagged protein in a bacterial culture, lyse the cells, isolate the affinity-tagged protein along with interacting proteins, and identify the affinity isolated proteins via mass spectrometry and informatics analysis. The pipeline was designed to be applicable to a wide array of gram negative bacterial species, and thus is sufficiently general to enable studies of any number of organisms that are of importance for DOE energy and environment missions. The cloning component of the pipeline is based on a flexible system (Gateway) that further expands the generality of the approach by allowing facile introduction of a wide variety of affinity (or other) tags.

The CMCS has made considerable progress toward using as affinity-tagged "baits" some 1200 *R. palustris* proteins that meet the following criteria: (1) the protein is predicted to be soluble, (2) the protein has been previously detected by mass spectrometry in proteomics studies. The results of our PPI survey in *R. palustris* are available via the Microbial Protein-Protein Interaction Database (MiPPI.ornl.gov). Statistical tools allow evaluation of the PPI based on characteristics of the data, and bioinformatics tools provide insights based on comparison of CMCS results to those from other techniques (e.g. gene expression measurements, PPI predictions) as well as PPI data from other organisms.

The results from the CMCS PPI pipeline are proving to be useful as a source of hypotheses for more detailed experiments aimed at particular pathways or systems in microbes. One example evolves around interactions observed in nitrogen fixing cells among proteins which are potentially involved in electron transfer to nitrogenase. Collaborative experiments involving the CMCS and the Harwood laboratory at the University of Washington are building on this result to explore the implications for production of hydrogen via the nitrogen fixation reaction. A further example involves study of a stress response pathway in *R. palustris* based on observed PPI involving proteins encoded by an operon that includes an ECF sigma factor, a putative response regulator, a putative histidine kinase, and an unknown protein.

Ongoing research in the CMCS is aimed at improving the throughput, applicability, and reliability of the PPI pipeline. Final validation and implementation of a robot-based protocol for affinity isolation will be completed in January 2008, removing a major bottleneck from the pipeline. Expansion of the CMCS pipeline to include membrane-associated proteins is underway. With these and other advances, the CMCS provides a unique resource for characterizing protein "machines" for the Genomics:GTL program. Details of these studies and other CMCS activities are covered in additional abstracts.

126

GTL

Advanced Data Analysis Pipeline for Determination of Protein Complexes and Interaction Networks at the Genomics:GTL Center for Molecular and Cellular Systems

Kevin K. Anderson,[2]* William R. Cannon,[2] Don S. Daly,[2] Brian S. Hooker,[2] Jason E. McDermott,[2] Gregory B. Hurst,[1] W. Hayes McDonald,[1] Dale A. Pelletier,[1] Denise D. Schmoyer,[1] Jenny L. Morrell-Falvey,[1] Mitchel J. Doktycz,[1] Sheryl A. Martin,[1] Mudita Singhal,[2] Ronald C. Taylor,[2] H. Steven Wiley,[2] and **Michelle V. Buchanan**[1] (buchananmv@ornl.gov)

[1]Oak Ridge National Laboratory, Oak Ridge Tennessee and [2]Pacific Northwest National Laboratory, Richland Washington

Project Goals: See goals for abstract 126.

The Genomics:GTL Center for Molecular and Cellular Systems (CMCS) is a DOE Center whose mission is to determine protein complexes and interaction networks for microbial systems. The CMCS is currently focusing on the completion of the characterization of soluble protein-protein interactions in *Rhodopseudomas palustris*. The CMCS approach combines expression of affinity tagged proteins, affinity purification of interacting proteins, and tandem mass spectrometric identification of these proteins. Our goal is to provide a capability for generating high quality protein-protein interaction data from a variety of energy- and environment-relevant microbial species. This poster provides a status report of the CMCS measurements of protein-protein interactions in *R. palustris*, which is of high relevance to DOE missions due to its ability to produce hydrogen, to degrade lignin monomers, and for its exceptional metabolic versatility. A critical component of the approach is our evolving data analysis pipeline.

As of early December 2007, nearly 1200 *R. palustris* genes have been cloned as Gateway entry vectors, and approximately 1060 expression clones for a dual affinity tag (6-His/V5) have been produced. Some 467 affinity-tagged bait proteins have been expressed, affinity purified,

* Presenting author

and subjected to mass spectrometry (MS) analysis to identify interacting proteins. Approximately 30% of these bait proteins are annotated as conserved hypothetical, conserved unknown, or unknown proteins.

The data pipeline for analysis of the data begins with a Laboratory Information Management System (LIMS) to capture the MS/MS data and descriptions regarding the biological and assay conditions (metadata). The LIMS maintains a detailed history for each sample by capturing processing parameters, protocols, stocks, tests and analytical results for the complete life cycle of the sample.

The resulting lists of potentially interacting prey proteins identified from MS/MS are statistically analyzed within a software environment specifically designed for working with biological networks. Bayes estimates of the confidence of the inferred associations are estimated for each bait/prey pair. For high confidence interactions, robust networks of interacting proteins are determined from patterns of interactions. The resulting protein networks are captured in a database within a publically accessible software environment (https://www.emsl.pnl.gov/SEBINI/). Using an exploratory data analysis tool that enables integration and analysis of interactions evidence obtained from multiple sources (CABIN, www.sysbio.org/capabilities/compbio/cabin.stm), the information on the nodes (proteins) and edges (interactions) can be linked to external and internal bioinformatic data. The internal bioinformatic data contains information on interologues derived from the *Bioverse* system, which provides additional information on protein interactions. The joint analysis of experimental data and multiple sources of bioinformatic data is done graphically through collective analysis of biological interaction networks (*Cabin*), a plug-in for the *Cytoscape* network visualization program.

These protein-protein interactions are disseminated through the publicly accessible Microbial Protein-Protein Interaction Database (MiPPI.ornl.gov). MiPPI is updated every 6 months (May and November). MiPPI provides tables of observed protein-protein interactions, as well as background information on CMCS measurement and analysis techniques. Various results (mass spectrometry results, corresponding metadata, and identified protein-protein interactions, including the statistical analysis scores) are also available for download in various file formats.

127

GTL

Analysis of the Dynamical Modular Structure of *Rhodopseudomonas palustris* Based on Global Analysis of Protein-Protein Interactions

William R. Cannon[2]* (william.cannon@pnl.gov), Mudita Singhal,[2] Ronald C. Taylor,[2] Don S. Daly,[2] Dale A. Pelletier,[1] Gregory B. Hurst,[1] Denise D. Schmoyer,[1] Jennifer L. Morrell-Falvey,[1] Brian S. Hooker,[2] W. Hayes McDonald,[1] **Michelle V. Buchanan**,[1] and H. Steven Wiley[2]

[1]Oak Ridge National Laboratory, Oak Ridge Tennessee; and [2]Pacific Northwest National Laboratory, Richland Washington

Project Goals: See goals for abstract 126.

Global determination of protein-protein interactions for *Rhodopseudomonas palustris* is the current target for the Genomics:GTL Center for Molecular and Cellular Systems (CMCS). *R. palustris* is a metabolically versatile anoxygenic phototrophic bacterium, and analyses have focused on protein interactions observed under differing conditions for nitrogen metabolism in which either NH_4^+ (fixed nitrogen) or N_2 serve as the primary source of nitrogen.

We have used the set of protein-protein interactions as the foundation for determining the dynamic modular structure of *R. palustris* regulatory networks. Global interactions determined by our affinity isolation pipeline are parsed into functional subnetworks by combining operon membership, gene regulatory information, gene expression information, phylogenetic profiling, gene neighborhood analyses and predicted interactions. Approximately 6,000 interactions between over 700 proteins were parsed in modular subnetworks and compared to the pattern of regulated gene expression observed under conditions of hydrogen utilization. We have also compared these functional modules with those inferred from protein interaction data gathered in other bacteria, such as *E. coli*. Our analysis indicates that different technologies for evaluating protein interaction networks have distinct inherent biases and that combining multiple data sources are likely to produce the most robust results. The subnetworks inferred from multiple data sources can provide novel hypotheses relating to previously unknown proteins and can serve as a foundation for further investigations—see the posters *Protein-Protein Interactions Involved in electron transfer to nitrogenase for Hydrogen Production in Rhodopseudomonas palustris* and

* Presenting author

poster *Identificataion of a Putative Stress Response Pathway and Novel Extracytoplasmic Function σ/Anti-σ Factors in the Anoxygenic Phototrophic Bacterium Rhodopseudomonas palustris by Protein-Protein Interactions* for detailed discussions of biological phenomena.

128

GTL

Characterization of a Stress Response Pathway in the Anoxygenic Phototrophic Bacterium *Rhodopseudomonas palustris*

Michael S. Allen[1]* (allenms@ornl.gov), Dale A. Pelletier,[1] Gregory B. Hurst,[1] Linda J. Foote,[1] Trish K. Lankford,[1] Catherine K. McKeown,[1] Tse-Yuan S. Lu,[1] Elizabeth T. Owens,[1] Denise D. Schmoyer,[1] Jennifer L. Morrell-Falvey,[1] W. Hayes McDonald,[1] Mitchel J. Doktycz,[1] Brian S. Hooker,[2] William R. Cannon,[2] and **Michelle V. Buchanan**[1]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee; and [2]Pacific Northwest National Laboratory, Richland, Washington

Project Goals: See goals for abstract 126.

Rhodopseudomonas palustris is an anoxygenic phototrophic bacterium possessing high metabolic diversity. As part of the Genomics:GTL Center for Molecular and Cellular Systems (CMCS) effort, this organism has been investigated for its ability to produce nitrogenase-mediated biohydrogen and its potential for bioremediation. Analysis of cytoplasmic protein fractions by shotgun proteomics has revealed several proteins up-regulated during growth on benzoate as well as under diazotrophic conditions. Among those was the putative extracytoplasmic function (ECF) σ-factor RPA4225. Subsequent large-scale protein-protein interaction experiments also revealed an interaction between the unknown protein RPA4224 and the putative response regulator RPA4223. RPA4224 and RPA4225 form a single operon in *R. palustris*, suggesting that this unknown protein may serve as an anti-σ factor. Organization of this operon along with the preceding response regulator gene RPA4223 is conserved among several α-Proteobacteria including *Sinorhizobium meliloti*, where the components have been shown to act as mediators of the global stress response. Additionally, we have found that the genomic location of the downstream gene RPA4226, a putative histidine kinase containing a predicted transmembrane domain, is also conserved among these bacteria. This suggests a potential role in the sensing and signal transduction of the stress response. These data underscore the utility of high-throughput methodologies to interrogate complex, multi-component systems and for generating new hypotheses regarding proteins about which little or nothing is known.

129

GTL

Protein-Protein Interactions Involved in Electron Transfer to Nitrogenase for Hydrogen Production in *Rhodopseudomonas palustris*

Dale A. Pelletier[1]* (pelletierda@ornl.gov), Erin Heiniger,[3] Gregory B. Hurst,[1] Trish K. Lankford,[1] Catherine K. McKeown,[1] Tse-Yuan S. Lu,[1] Elizabeth T. Owens,[1] Denise D. Schmoyer,[1] Jennifer L. Morrell-Falvey,[1] Brian S. Hooker,[2] W. Hayes McDonald,[1] Mitchel J. Doktycz,[1] William R. Cannon,[2] Caroline S. Harwood,[3] and **Michelle V. Buchanan**[1]

[1]Oak Ridge National Laboratory, Oak Ridge Tennessee; [2]Pacific Northwest National Laboratory, Richland, Washington; and [3]University of Washington, Seattle, Washington

Project Goals: See goals for abstract 126.

The goal of the Center for Molecular and Cellular Systems (CMCS) is to identify protein-protein interaction networks that form the molecular basis of biological function in bacterial species relevant to the Genomics:GTL program. *Rhodopseudomonas palustris* is a metabolically versatile anoxygenic phototrophic bacterium that is emerging as a model system for nitrogenase-mediated biohydrogen production. This process requires the integration of several metabolic and regulatory networks, including nitrogen metabolism, photosynthesis and carbon metabolism. Although the nitrogenase enzyme has been the focus of much research, we have a poor understanding of the organization of cellular components facilitating the flow of electrons derived from carbon metabolism to nitrogenase in *R. palustris* and other diazotrophic bacteria. To better understand this and other processes, we have begun mapping the protein-protein interactions in photoheterotrophically grown *R. palustris.* Shotgun proteomics and microarray analysis have identified proteins that are upregulated in *R. palustris* cells grown in the absence of fixed nitrogen. These proteins where subsequently analyzed to identify protein-protein interactions by affinity isolation and mass spectrometry. This analysis revealed interactions among numerous proteins including FixABCX, a predicted protein complex hypothesized to have a role in transfer

* Presenting author

of electrons to nitrogenase. Subsequently we found that a *fixABCX* mutant was deficient but not completely blocked in its ability to grow under nitrogen fixing conditions. This mutant was also deficient in nitrogenase activity. Supplying fixABCX in trans restored the growth phenotype. RPA1927 and RPA1928 encode proteins of unknown function that are also highly expressed under nitrogen-fixing growth conditions. While the functions of RPA1927 and RPA1928 are unknown, the presence of a predicted ferredoxin-like iron-sulfur cluster in RPA1928 implicates this protein in electron transfer. Additionally a novel putative interaction was identified between the proteins encoded by RPA1927 and RPA1928 and the FixABCX complex implying a potential role in electron transfer. An RPA1927-RPA1928 deletion strain has been constructed and growth phenotypes are under investigation. These studies have increased our understanding of the pathways and protein-protein interactions that occur in *R. palustris* cells grown under nitrogen-fixing and hydrogen producing conditions. These results as well as the results of future interaction studies will allow for modeling and metabolic engineering of this organism for increased yields of biological hydrogen.

130

GTL

Application and Optimization of a Multi-Use Affinity Probe (MAP) Toolkit for Systems Biology

M. Uljana Mayer, Baowei Chen, Yijia Xiong, and **Thomas C. Squier*** (thomas.squier@pnl.gov)

Pacific Northwest National Laboratory, Richland, Washington

Project Goals: Newly synthesized MAPs built upon the cyanine dyes used for single molecule imaging offer the potential for multicolor measurements of protein localization and associations, and provide a path-forward for the high-throughput parallel characterization of protein-protein networks using a single tagging step and affinity technology. Because protein complexes can be released using simple reducing agents, low-affinity binding interactions can be captured, identified, and validated using the same MAPs. However, the robust utilization of MAPs requires the development of standard protocols that provide recipes and outline limitations regarding how MAPs can be used to image and purify protein complexes. We will focus on the application of existing MAPs and associated resins that we have synthesized, paying particular attention to the following deliverables. I. Demonstrate Utility of New Brighter MAPs (i.e., AsCy3) to Image Bacterial Proteins. II. Establish Ability of MAPs to Isolate Protein Complexes in Comparison with Established Tandem Affinity Purification Approaches. III. Benchmark Requirements of MAPs for Imaging and Protein Complex Measurements.

Summary: Newly synthesized multiuse affinity probes (MAPs) built upon the cyanine dyes used for single molecule imaging permit multicolor measurements of protein localization and associations, and provide a path-forward for the high-throughput parallel characterization of protein-protein networks using a single tagging step and affinity technology. Genetically encoded tags are engineered onto proteins of interest, and subsequently labeled using MAPs. The small tag size and the ability to use MAPs to image proteins under anaerobic conditions has substantial advantages relative to other technologies involving, for example, fluorescent proteins whose large size, slow folding kinetics, and requirement for molecular oxygen for chromophore biosynthesis prevent their robust application in a range of bacterial systems. Following the immobilization of MAPs on solid supports, protein complexes can be released using simple reducing agents. Low-affinity binding interactions are readily captured, identified, and validated using the same MAPs. We report the application of existing MAPs and associated resins that we have synthesized, providing examples of the following applications using the cytosolic RNA polymerase (RNAP) complex and integral membrane proteins associated with metal reduction in *Shewanella*.

I. *Demonstrated Utility of MAPs to Image Bacterial Proteins*. Live cell imaging has demonstrated the utility of using MAPs to label both cytosolic and membrane proteins in highly pigmented bacteria. The development of red/infrared MAPs provides an effective means of monitoring protein-protein interactions in highly pigmented microbes where the interference with existing chromophores interferes with the utilization of previously developed probes. Further, MAPs permit the labeling of proteins following their cellular localization, and coupled with their small size permit pulse-chase measurements of cellular trafficking.

II. *Established Ability of MAPs to Isolate Protein Complexes in Comparison with Established Tandem Affinity Purification Approaches*. Following immobilization of MAPs on solid supports, intact supramolecular protein complexes are eluted using mild reducing conditions for protein identification using mass spectrometry. In comparison with traditional tandem affinity approaches, numerous low-affinity binding interactions are retained. As the

* Presenting author

intact and functional complex is eluted, complementary structural and functional measurements are possible to assess the consequences of macromolecular organization without the need to for complex reconstitution experiments of purified proteins.

III. *Use of MAPS for the Validation of Protein-Protein Interactions and Measurements of Catalytically Important Motions for High-throughput Functional Screens*. Multicolor measurements permit the facile validation of protein-protein interactions and structural arrangements within individual protein complexes in either cellular lysates or living cells. Using defined model systems, we demonstrate the ability to identify binding interactions between proteins in complex using both fluorescence correlation spectroscopy and energy transfer measurements. Complementary measurements demonstrate the ability to assess functional protein motions that offer a means to assess changes in protein function in living cells in response to changes in environmental conditions.

131

GTL

Development of Highly Efficient Bacterial Hosts for High Throughput Recombinant Membrane Protein Production

Hiep-Hoa T. Nguyen* (hiephoa@its.caltech.edu), Sanjay Jayachandran, and Randall M. Story

TransMembrane Biosciences, Pasadena, California

Project Goals: The objective of this project is to develop superior host strains for efficient and high yield production of correctly folded and functional recombinant membrane proteins.

Membrane proteins and enzymes (for instance channels, receptors, and transporters) are involved in critical cellular processes, but our understanding of these important biological molecules at molecular levels falls behind those of soluble proteins. For example, while thousands of X-ray crystal structures of soluble proteins are known, only ~80 structures of unique membrane proteins are currently available. This disparity is increasing with the surge in data generated by various genome and structural genomics projects. The most significant problem precluding any structural characterization of membrane proteins is their low levels of biosynthesis. As a result, obtaining significant quantities (even at milligrams scale) of purified membrane proteins for biochemical and biophysical studies has been a major obstacle. An effective membrane protein overexpression system would be indispensable and allow us to tackle this intractable but very significant problem in membrane protein biochemistry. All of the commercially available protein expression vehicles yield very poor results for membrane proteins even though the most powerful systems can generate *gram* quantities of recombinant soluble proteins even with small scale fermentation. Considering the ineffectiveness of current expression systems in overexpressing membrane proteins, it is clear that economical and effective expression systems for membrane proteins are needed. The objective of this proposal is to develop high yield membrane protein expression systems. Specifically, we seek to obtain superior bacterial hosts for membrane protein production. The goal of this research is to improve further the yield of recombinant membrane proteins produced by our bacterial hosts to the extent that significant quantities can be obtained with small scale cultures. We have pursued this objective through a combination of rational genetic engineering and directed evolution/screening process to obtain superior membrane protein production hosts using *Escherichia coli*, a familiar, robust, and highly amendable microorganism. We have been working in creating strains that overexpress proteins involving in membrane protein biosynthesis using background mutants obtained through screening efforts. The background mutants were selected through screening assays that demonstrate the mutants' capability to tolerate high expression levels of recombinant membrane proteins. Evaluation of the improved strains obtained through genetic engineering by test-expression of a number of model recombinant membrane proteins is in the way. As a demonstration for the power of our expression technologies, recently, we produced >100 mg of purified wild type and Se-Met labeled Rh protein (a channel with 11 transmembrane helices) within 2 months and solved the X-ray crystal structure of this channel in 5 months. The 1.8 A resolution X-ray crystal structure of this channel and related results were recently published in *PNAS*. The speed and efficiency of this effort is a validation for our approach in membrane protein production.

* Presenting author

Validation of Genome Sequence Annotation

132

GTL

A High Thoughput Proteomic and Protein Expression Strategy for Annotation of Fungal Glycosyl Hydrolases

Scott E. Baker[1]* (scott.baker@pnl.gov), Jon K. Magnuson,[1] Ellen A. Panisko,[1] Adrian Tsang,[2] and Frank Collart[3]

[1]Fungal Biotechnology Team, Energy and Environment Directorate, Pacific Northwest National Laboratory, Richland, Washington; [2]Fungal Genomics, Concordia University, Montreal, Canada; and [3]Biosciences Division, Argonne National Laboratory, Lemont, Illinois

Project Goals: The objective of this project is to use high throughput proteomics, protein expression and enzyme assays to rapidly generate functional data and annotation for secreted proteins, specifically glycosyl hydrolases.

There has been an exponential increase in the number of microbial genomes sequenced but not the number of methods for functional annotation. These still rely primarily on sequence similarity whether performed by algorithms or by manual curators. This is very effective with enzymes that have been well characterized and exhibit a high degree of sequence identity across phyla, such as central metabolic enzymes. However, when insufficient experimental information on substrate specificity is available, only general functions can be assigned to the genes, or worse, improper annotations can be made and quickly propagated through the databases. This is especially true for glycosyl hydrolases as groups of families based on sequence similarities often have a variety of substrate specificities. The crucial role of fungi in environmental carbon and nitrogen cycling stems from their absorptive nutritional strategy. The glycosyl hydrolases that fungi secrete are integral to this strategy and in addition, comprise a large fraction of fungal genes. Filamentous fungal genomes contain between 100 and 220 glycosyl hydrolase genes for the breakdown and utilization of plant, fungal and prokaryotic cell walls. Glycosyl hydrolases (GH) are critical for the hydrolysis of plant derived biomass for subsequent fermentation to liquid fuels such as ethanol as well as products such as organic acids. There are over 150 recognized GH activities in the enzyme classification system (EC 3.2.1._), and this category is growing. The CAZy database (http://afmb.cnrs-mrs.fr/CAZY/) is an excellent source of information about glycosyl hydrolases (and other carbohydrate active enzymes). The GHs are currently divided into over 100 families based on sequence similarity and these families are extremely helpful for annotation of genomes. However, even within families, multiple activities are often found, such that primary sequence homology alone is insufficient to definitively assign a function. Therefore, we are developing a high throughput protein-centric annotation pipeline for fungal GHs (Figure) consisting of mass spectrometry based proteomic identification of secreted proteins from fungi grown on a variety of biomass substrates, high throughput protein expression of fungal GHs prioritized by the proteomic analyses, and substrate development for multiplexed assay of native and recombinant GHs. A variety of substrates and assay methods will be developed for high-throughput functional classification and characterization of fungal secreted enzymes. The focus will be on known and potentially novel glycosyl hydrolase families.

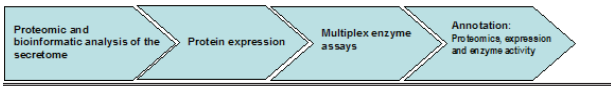


Figure. Pipeline for high throughput functional annotation of fungal glycosyl hydrolases.

133

GTL

Assignment of Enzymatic Function for Core Metabolic Enzymes

Vincent Lu,[1]* Gopi Podila,[2] Michael Proudfoot,[3] Alexander Yakunin,[3] and **Frank Collart**[1] (fcollart@anl.gov)

[1]Biosciences Division, Argonne National Laboratory, Lemont, Illinois; [2]Department of Biological Sciences, University of Alabama, Huntsville, Alabama; and [3]Banting and Best Department of Medical Research, University of Toronto, Toronto, Canada

Project Goals: 1. Functionally map the set of conserved hypothetical proteins from *Shewanella oneidensis* which contains ~800 members (TIGR annotation). The rationale for screening the set of conserved hypothetical from a single genome is to benchmark the utility of the enzymatic screening approach for the improvement

* Presenting author

of functional assignment for a large set of proteins of unknown function. 2. The second aim will be to apply a directed screening strategy to uncharacterized proteins of the haloacid dehalogenase (HAD)-like hydrolase superfamily, which will be tested for the presence of dehalogenase, phosphonatase, phosphatase, or glucomutase activities for HAD-like hydrolases. This aim will use several strains of *Shewanella* and the symbiotic fungus *Laccaria bicolor*. This objective will provide a foundation to assess the capabilities for specific functional assignments for a substantial number of unknown prokaryotic and eukaryotic proteins.

With over 600 genomes with complete sequences currently available in public databases and thousands of genome sequence projects in progress, there's a pressing need to effectively annotate genomic sequences quickly and accurately for functional activity. The main objective of this proposal is to experimentally annotate (assign a biochemical function) a large group of conserved hypothetical proteins using high throughput protein production and enzymatic screening methods. This approach for experimental annotation will be applied to hypothetical proteins from a prokaryote and a eukaryote of programmatic interest. In the first stage of the project we will functionally map the set of conserved hypothetical proteins from *Shewanella oneidensis* which contains ~800 members (Figure). The rationale for screening the set of conserved hypothetical from a single genome is to benchmark the utility of the enzymatic screening approach for the improvement of functional assignment for a large set of proteins of unknown function. A second component of the project will be to apply a directed screening strategy to uncharacterized proteins of the haloacid dehalogenase (HAD)-like hydrolase superfamily, which will be tested for the presence of dehalogenase, phosphonatase, phosphatase, or glucomutase activities for HAD-like hydrolases. This aim will use several strains of *Shewanella* and the symbiotic fungus *Laccaria bicolor*. Targets from *L. bicolor* will be amplified from cDNA clones, clone libraries or generated using a PCR-based gene synthesis approach. This objective will provide a foundation to assess the capabilities for specific functional assignments for a substantial number of unknown prokaryotic and eukaryotic proteins.

For protein production, we will use the efficiency of automated strategy to implement a parallel pipeline consisting of an *E. coli* and yeast expression systems. The screening strategy uses a tiered approach where targets are categorized using a series of general screens and then rescreened for specific functional assignments using a directed series of natural substrates (Figure). The general screening assays have relaxed substrate specificity and are designed to identify the subclass or sub-subclasses of enzymes (phosphatase, phosphodiesterase/nuclease, protease, esterase, dehydrogenase, and oxidase) to which the unknown protein belongs. Further biochemical characterization of proteins can be facilitated by the application of secondary screens with natural substrates (substrate profiling).

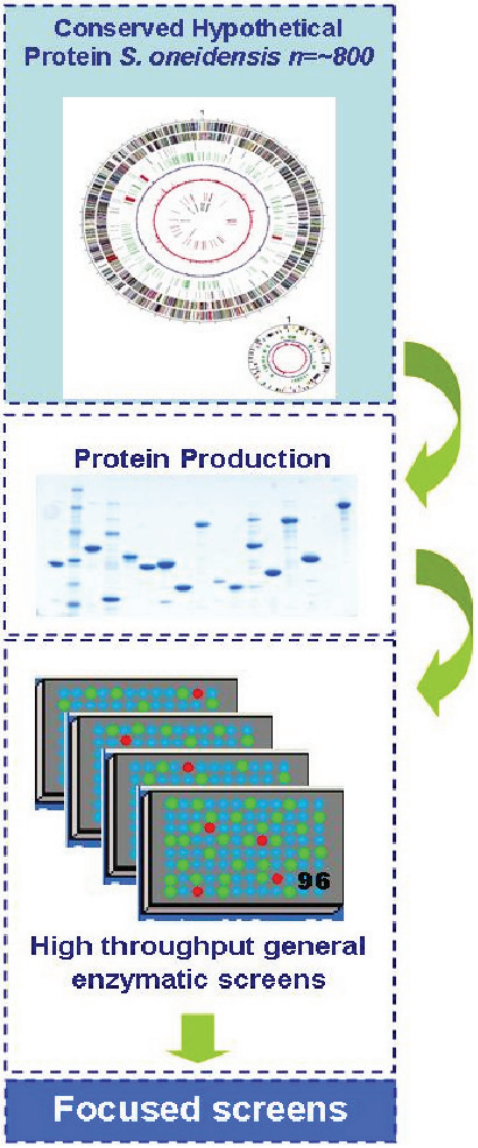


Figure. Illustration of the protein production and enzymatic screening process.

* Presenting author

134

GTL

Characterization of Sensor Proteins and Domains

Sarah Giuliani,[1]* Lori Field,[1] F. William Studier,[2] Lisa M. Miller,[3] and **Frank Collart**[1] (fcollart@anl.gov)

[1]Biosciences Division, Argonne National Laboratory, Lemont, Illinois; [2]Biology Division, Brookhaven National Laboratory, Upton, New York; and [3]National Synchrotron Light Source, Brookhaven National Laboratory, Upton, New York

Project Goals: 1. Specific functional assignments for several classes of sensor proteins. 2. Functional assignments for a set of homologs that can be used to define sequence motifs that can improve annotation assignments based on sequence alignments. 3. An expression clone library with extensive characterization data. 4. A novel set of versatile vectors customized for expression of sensor proteins.

All cells contain proteins that sense the environment and mediate transport and signaling events that lead to changes in metabolism and/or initiate changes in gene expression at the level of transcription. Mapping of ligands with these binding/sensor proteins is critical to our understanding of cell biochemistry and is essential for modeling cellular processes and the rational design of engineered organisms. The goal of this project is to evaluate a series of experimental techniques to screen for potential ligands that bind to "sensor type" proteins. The experimental approach is based on the observation that ligand binding for many proteins can alter stability of the protein. A fluorescence-based thermal shift assay was used for the identification of bound ligands and assignment of function. This is a target independent assay that uses a florescent dye to monitor protein unfolding and has been widely used for the assessment of ligand binding. This assay uses a commercially available real-time PCR instrument where thermal melting curves of the protein/ligand combinations can be screened in a 96-well plate format. To illustrate the suitability of this approach for ligand binding, we used Carbonic Anhydrase I [CAI, (EC 4.2.1.1)] as a positive control test protein. Our analysis indicated that the native form of this protein displays a concentration-dependent thermal shift (Fig. 1a). Addition of the tight binding inhibitor trifluoromethanesulfonamide (TFMSA) results in an increase in the Tm indicating enhanced protein stability (Fig. 1b). This approach will be used to improve gene/protein functional assignments of sensor type proteins by developing high throughput methods to match ligands with their binding proteins. A library of preliminary candidate binding ligands was generated using structural models in the Protein Data Bank (PDB) to identify sensor type proteins containing bound ligands. Our results show there are many structures with unique combinations of sensor type proteins and ligands. The bound ligands can be grouped into several categories such as amino acids, metals, small ions, sugars, and vitamins. Protein sequences derived from the PDB set will be used to identify potential homologs in a set of reagent genomes as candidates for functional screening. Targets for screening will be produced at Argonne using an establish pipeline for protein production and will be extensively characterized as assurance of protein or domain structural integrity which is necessary prior to ligand screening. A parallel effort at Brookhaven National Lab led by Dr. Studier will test potential strategies for increasing the efficiency of producing soluble, functional proteins, which could then be integrated into the Argonne protein production pipeline. This project will provide specific functional assignments for an important class of regulatory molecules and generate an expression-clone resource for future structural and functional studies.

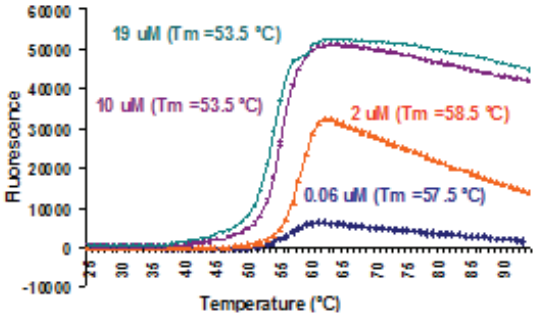


Fig. 1a. Protein concentration dependence of thermal shift assay with CAI and 5X SYPRO orange dye.

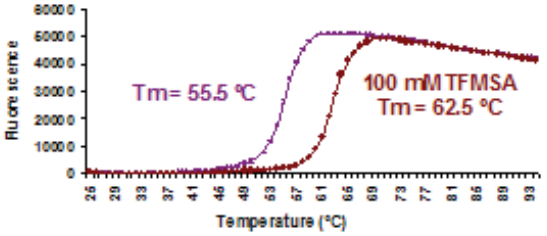


Fig. 1b. Thermal shift assay using 10 uM CAI and 5X SYPRO orange dye with and without ligand (100mM TFMSA)

* Presenting author

135

GTL

Protein Annotation from Interaction Networks using Zorch and Bayesian Functional Linkages

Richard Llewellyn[1]* and **David Eisenberg**[2] (david@mbi.ucla.edu)

[1]UCLA-DOE Institute for Genomics and Proteomics and [2]Department of Chemistry and Biochemistry, University of California, Los Angeles, California

Project Goals: Our goal is to provide a rigorous platform, known as Bayesian functional linkages, to integrate various types of functional inferences into a single description of the biological role of an uncharacterized target protein. The resulting annotation is a product of integrating the biological processes of the characterized functionally-linked proteins with any existing knowledge about the cellular location or molecular function of the target protein. We demonstrate its performance in yeast with links defined by zorch, a measure of connectivity of proteins in an interaction network (DIP). A central part of our goal is to identify and control sources of error in computational analysis so that as much information as possible can be used to infer the function of unknown proteins without diluting the accuracy of results. Bayesian functional linkages is a generalized method that we can extend to annotating proteins from uncharacterized prokaryotic genomes by combining inferences from homology with genome context and phylogenetic profiles.

Functional linkages describe relationships between proteins that work together to perform a biological task. Here we develop a general framework known as Bayesian functional linkages to annotate target proteins (Figure). We demonstrate its performance with links defined by zorch, a measure of connectivity of proteins in an interaction network. We started with the protein-protein interaction networks archived in the Database of Interacting Proteins (DIP).

Our goal is to provide a rigorous platform to integrate various types of functional inferences into a single description of the biological role of a protein. The resulting annotation is a product of integrating the biological processes of the linked predictor proteins with any existing knowledge about the cellular location or molecular function of the target protein. In Bayesian terminology, the data are the given annotations of the predictor proteins, the type of evidence supporting these annotations, and the strength of each functional linkage to the target. The hypotheses are the possible roles of the target protein as described by Gene Ontology biological process annotations. We address several of the common challenges of annotation by functional linkage: the contribution of each linked predictor protein is modulated by both the strength of its linkage to the target and the confidence that its characterized functions are correct according to their supporting evidence, in a manner that accounts for predictors with multiple annotations. The Bayesian likelihood focuses on the predictor annotations that most strongly support each hypothesis and quantifies the observation that different types of functional linkages are more likely to link proteins with particular types of functions. The contribution of any prior knowledge of the target's cellular location or molecular function is balanced against the number and quality of available functional linkages, with the final prediction given either as a posterior distribution over annotations that retains the influence of competing hypotheses, or as a single Bayes classifier of the most probable, general description of the biological process of the target.

We show that functional linkages quantified by zorch are good predictors of the biological process of proteins in *Saccharomyces cerevisiae*, and that the lack of high quality links can often be mitigated by the use of many weak ones, e.g. those inferred only from indirect or high-throughput protein-protein interactions, so that functional annotation can be extended to uncharacterized yeast proteins that lack reliably determined interacting partners.

A central part of our goal is to identify and control sources of error in computational analysis so that as much information as possible can be used to infer the function of unknown proteins without diluting the accuracy of results. We estimated the error in the existing GO annotations of predictor proteins when supported by different evidence types, such as Inferred from Expression Profile or Reviewed Computational Analysis, by comparing the accuracy rate of predicting known proteins annotated via gold standard evidence (i.e. Inferred from Direct Assay) from a linked protein with varying types of evidence. We also develop a novel method of treating GO annotations as hypotheses that explicitly addresses the incompleteness of any current ontology.

Bayesian functional linkages is a generalized method that we can extend to annotating proteins from uncharacterized prokaryotic genomes by combining inferences from homology with genome context and phylogenetic profiles. We expect that our focus on combining different types of inference while controlling sources of error will provide more accurate and higher resolution annotations

* Presenting author

than methods that treat functional linkages separately from homology.

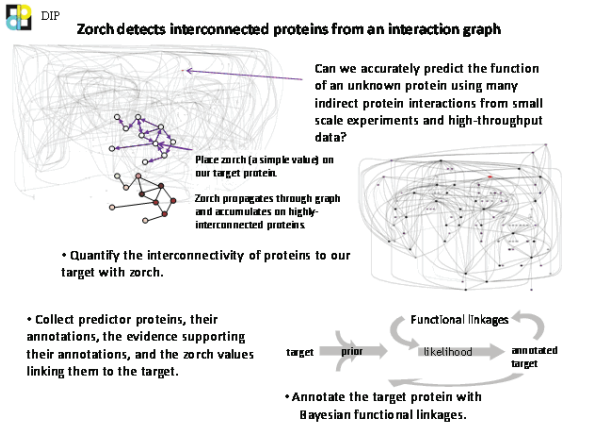


Figure. Overview of protein annotation with zorch and Bayesian functional linkages. Proteins interconnected to the unknown target are identified with zorch and integrated with any existing prior knowledge to predict the biological process.

We thank DOE BER for support.

136

GTL

Annotation of Novel Enzymatic Functions in Methanogens

Ethel Apolinario,[1] Zvi Kelman,[2] Jing Li,[3] Basil J. Nikolau,[3] Kevin Sowers,[1] and **John Orban**[2]* (orban@umbi.umd.edu)

[1]Center of Marine Biotechnology, University of Maryland Biotechnology Institute, Baltimore, Maryland; [2]Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, Maryland; and [3]W. M. Keck Metabolomics Research Laboratory, Iowa State University, Ames, Iowa

Project Goals: An integrated high-throughput approach is being developed to functionally annotate a large group of poorly understood genes in the methanogenic archaeon, *Methanosarcina acetivorans*. The focus is on genes predicted to encode enzymes, the substrate(s) and products of which are unknown. Approximately 2226 of the 4524 genes in *M. acetivorans* fall into this category and include genes possibly involved in processes such as methanogenesis, nitrogen fixation, and carbon assimilation. The biochemical functions of these putative enzymes will be accurately annotated using a combination of gene knockouts, high throughput metabolomic analysis with mass spectrometry (MS), automated screening of implicated metabolites with nuclear magnetic resonance spectroscopy (NMR), and biochemical assays.

Targets for study will include uncharacterized genes that have been associated with metabolic pathways by transcript expression using genomic microarrays. Also gene products with sequence or structural homology to proteins with a putative enzymatic function will be selected as targets. Gene disruption will be conducted by complementing proline auxotrophy in a *proC*- strain of *M. acetivorans*, which allows later complementation with a plasmid containing the wild type gene and puromycin resistance (pac) cassette. Genes will be disrupted by insertion of *proC* into the target open reading frame (ORF) followed by homologous recombination into a proline auxotroph of *M. acetivorans*. In cases where multiple gene disruptions are necessary, a directed markerless genetic disruption system will be used.

Using MS-based analyses, metabolic profiles will be compared between isogenic strains that carry either wild type or gene-knockout alleles at the locus of interest. Changes in metabolite pools in knockouts will provide clues about the specific pathway or set of pathways relevant to the function of the gene target. The extent to which MS data alone will infer a function will depend on whether there are any other intersecting pathways that may shunt elevated metabolites off in other directions.

To further increase the probability of obtaining useful functional annotations, we propose to screen potential substrates and products, or their structural analogs, from affected pathways for interaction with the putative enzyme in question using NMR methods that are not limited by molecular weight considerations. Metabolites and analogs will be obtained from commercial sources and proteins needed for screening will be expressed and purified in-house.

Compounds that interact with the protein of interest will be tested in a defined biochemical enzyme assay to validate the functional annotation. For example, if a putative enzyme is thought to be a methyltransferase then the compounds/metabolites that interact with it based on combined MS/NMR data will be used in an assay to determine whether a methyl group can be transferred to them. A range of assays will be utilized and MS and NMR will also be employed to monitor turnover reactions.

Preliminary results illustrating our general approach will be presented.

* Presenting author

137

GTL

Genemap-MS: High Throughput Mass Spectrometry Methods for Functional Genomics

Trent R. Northen[1,2]* (trnorthen@lbl.gov), Linh Hoang,[2] Steven M. Yannone,[1] Jason Raymond,[4] Jill Fuss,[1] Jinq-Chyi Lee,[3] Der-Ren Hwang,[3] Chi-Huey Wong,[3] John Tainer,[1,2] and **Gary Siuzdak**[1,2]

[1]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; [2]Department of Molecular Biology and Scripps Center for Mass Spectrometry and [3]Department of Chemistry and the Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, California; and [4]School of Natural Sciences, University of California, Merced, California

Project Goals: The utility of genetic information being derived from sequencing efforts is diminished by the incomplete and sometimes incorrect annotations associated with "completed" genomes. Homology-based protein function predictions are limited by evolutionary processes that result in conserved domains and sequence being shared by enzymes of widely diverse functions. Therefore, additional experimental datasets directed at validating and improving genome annotations are required. Here an integrated approach is used to develop universally applicable high-throughput (HT) methods for validating genome annotation using mass spectrometry (MS) based proteomics, metabolomics, and our developing technologies for detecting biochemical activities on arrayed metabolite substrates. This leverages expertise with developing and applying MS technologies to generate datasets directly applicable to validation. These general and multifaceted approaches focus on: 1) protein expression 2) metabolic pathways and 3) biochemical activities. The datasets from these analyses are integrated into computational metabolic networks to provide physical and functional validation of the many hypothetical and predicted proteins and activities in current genome annotations. We provide a balance of mature and robust MS technologies with new surface based MS technologies and the expansion of our METLIN metabolite database with the ultimate goal of addressing specific DOE needs for exploiting microbial processes for bioenergy production.

Mass spectrometry's ability to efficiently generate intact biomolecular ions in the gas phase has led to its widespread applications including metabolomics, proteomics, and biological imaging. Matrix-Assisted Laser Desorption/Ionization (MALDI) and Electrospray Ionization (ESI) have been at the forefront of these developments. We recently introduced Nanostructure-Initiator Mass Spectrometry (NIMS), a sensitive new tool for spatially defined mass analysis which complements existing methods by enabling the analysis of metabolites from tissues, cells, microarrays, and etc. with high sensitivity. NIMS utilizes 'initiator' molecules trapped in nanostructured surfaces or 'clathrates' to release/ionize intact molecules adsorbed from the surface. This technology has recently been extended for the direct screening of cell lysates and environmental samples for enzymatic activities at high temperatures and low pH values. Using this approach a new thermophilic galactosidase was identified from a Yellowstone hot springs microbial community. In addition the optimal pH, temperature, and enzyme inhibition were screened *in situ.* This general approach provides an efficient method for screening environmental sample prior to sequencing and cloning efforts and without obtaining pure cultures.

138

GTL

The Application of Phage Display to Advanced Genome Annotation: *C. Thermocellum* as an Example

Andrew Bradbury* (amb@lanl.gov)

Los Alamos National Laboratory, Los Alamos, New Mexico

Folding reporters are proteins with easily identifiable phenotypes, such as antibiotic resistance or fluorescence, whose folding and function is compromised when fused to poorly folding proteins or random open reading frames. We have found that when DNA fragments are fused to a ß lactamase folding reporter, selection for fragments of real genes, as opposed to random ORFs, tends to occur. *We hypothesize that folding reporters can be used on a genomic scale to select collections of correctly folded protein domains from the coding portion of the DNA of any organism. This technology will be applicable to any intronless genome, or collection of open reading frames, without extensive analysis or primer synthesis. This can be considered to be the protein equivalent of shotgun sequencing, and could be termed the "domainome", to extend an overused cliché.* It is expected that the protein fragments obtained by this approach will be well expressed and soluble, making them suitable for structural studies, antibody generation, protein/substrate binding analyses, domain shuffling for enzyme evolution and protein chips.

* Presenting author

Of the many functions undertaken by modular domains, molecular recognition is the most easily assessed, and the most straightforward to translate into gene annotation. By cloning the "domainome" directly in a phage display context, it will be possible to select gene fragments encoding domains with specific binding properties (e.g. to other proteins, domains, metabolites, enzyme substrates), *providing essential experimental information for gene annotation.* We expect this concept can be extended to activity based probes (ABPs)[1], to identify domains with catalytic activities. Once a domain fragment library has been created, it is a renewable resource, easily retested against new potential binding partners or ABPs.

This hypothesis will be tested with DNA from the genome of *C. thermocellum* and specifically applied to the identification of cellulose binding domains (CBDs). This genome is such a rich source of cellulase genes that it provides an excellent model system for the proposal described here, with the possibility of experimentally identifying novel cellulases or other enzymatic activities linked to CBDs, and subsequently extending it to the analysis of cohesions and dockerins.

In this poster the concept, and the progress we have made with the display of cellulose binding domains derived from *C. thermocellum* will be described.

139

GTL

Phylogenomics-Guided Validation of Function for Conserved Unknown Genes

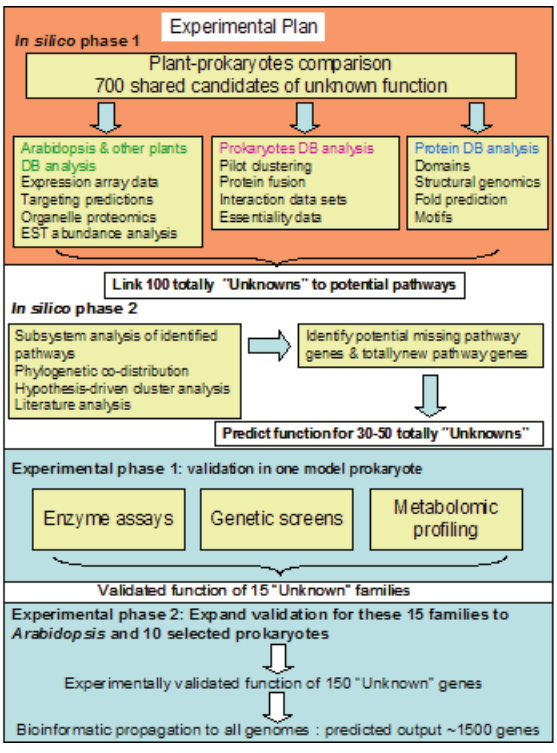
Valérie de Crécy-Lagard[1]* (vcrecy@ufl.edu), Basma El Yacoubi,[1] Crysten Haas,[1] Valeria Naponelli,[2] Alexandre Noiriel,[2] Jeffrey C. Waller,[2] and Andrew D. Hanson[2]

[1]Department of Microbiology and Cell Science and [2]Department of Horticultural Sciences, University of Florida, Gainesville, Florida

Project Goals: Identifying the function of every gene in all sequenced organisms is a major challenge of the post-genomic era. Our objective is to use an integrative approach to predict and experimentally verify the in-vivo function of genes that lack homologs of known function ('unknown' gene families) and that are highly conserved among prokaryotes and plants.

The approach has four phases. A phylogenomic analysis comprising two *in silico* phases will lead to prediction of function for 30-50 unknown gene families. Then come two experimental phases: validation of the prediction in one prokaryote and *Arabidopsis*, then extension of the validations to ten other organisms using phenotypic and enzyme assays developed in the previous phase. We expect the final outcome to be experimentally validated functions for 15 families of unknowns, which translates into ~150 individual genes. These functions can then be propagated with confidence to all genomes, leading to the functional annotation of an estimated 1500 genes.

The phylogenomic analysis is ongoing, but in pilot work we have predicted the general function of around ten families. Further bioinformatic analysis combined with experimental validations has led to more precise functional prediction for four of these families. These will be presented in more detail and consist of: 1) The universal YrdC/Sua5 family (COG009) that is involved in the modification of tRNA; 2) The CobW family (COG0523) that is a metal chaperone that might have a important role in zinc homeostasis. 3) YgfZ, a folate binding protein that could be involved in repair of iron sulfur proteins; 4) The pterin carbinolamine dehydratase family (COG2154), which occurs in many organisms that lack the pterin recycling pathway requiring the dehydratase and that could have another function in molybdenum cofactor maintenance.



* Presenting author

140

GTL

Prodigal: A New Prokaryotic Gene Identification Program with Enhanced Translation Initiation Site (TIS) Prediction

Doug Hyatt, **Loren Hauser*** (hauserlj@ornl.gov), Frank Larimer, and Miriam Land

Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee

Project Goals: Enhanced quality of gene prediction and annotation in JGI microbial organisms.

High-quality annotation of microbial genomes remains an ongoing challenge for the Joint Genome Institute. For the past several years, Oak Ridge National Laboratory (ORNL) has aided in gene prediction and functional analysis of numerous microbial organisms. As a result of the comprehensive review and curation of a large number of genomes ranging from low-GC to high-GC and from bacteria to archaea, numerous areas of improvement have been found. Predicting the correct number of genes, reducing the number of false positives, correctly locating short and laterally transferred genes, performing robustly in high-GC-content genomes, and accurately finding the translation initiation site (TIS) of genes continue to be challenges that will enhance the quality of the final JGI annotations submitted to Genbank and placed in IMG.

Prodigal (Prokaryotic Dynamic Programming Genefinding Algorithm) was developed at ORNL to address many of the "real-world" challenges discovered through many hours of manual curation of microbial genomes. In particular, the previous pipeline (based on the genefinders Critica and Glimmer) often lengthened genes in high-GC content genomes and incorrectly omitted genes that would overlap the erroneously long genes. We decided to address this issue with a new gene identification algorithm that would perform robustly in high-GC content genomes. Prodigal's self-training methodology is based on a detailed analysis of the GC-frame-plot of the organism in question. The training process consists of determining the statistical significance of G and C in different frame positions and performing a dynamic programming algorithm using this information to construct an initial training set of genes. This is a novel approach compared to other programs, which construct their training sets based merely on all open reading frames (ORFs) above a particular length. Our implementation of a coding scoring function based on this training set was found to perform well in both low-GC and high-GC genomes.

The other improvement to the annotation pipeline is improved start site prediction. Prodigal contains a novel method for examining the upstream regions for ribosomal binding site (Shine-Dalgarno) motifs. The statistical significance of various motifs relative to the background is determined automatically by an iterative algorithm which learns the organism's preference for various RBS motifs. Results for this enhanced start site prediction are presented, as well as overall results for locating the 3' end of genes. In addition, future improvements to the algorithm are discussed, such as validation through proteomics data and improvement of start site prediction via signal peptide information.

* Presenting author

Computing Resources and Databases

141

GTL

Further Refinement and Deployment of the SOSCC Algorithm as a Web Service for Automated Classification and Identification of *Bacteria* and *Archaea*

J. Fish,[1] Q. Wang,[1] S.H. Harrison,[1] T.G. Lilburn,[2] P.R. Saxman,[3] J.R. Cole,[1] and **G.M. Garrity**[1]* (garrity@msu.edu)

[1]Michigan State University, East Lansing, Michigan; [2]American Type Culture Collection, Manassas, Virginia; and [3]University of Michigan, Ann Arbor, Michigan

Previously, we had demonstrated that techniques such as principal components analysis (PCA), could be useful in unraveling discontinuities between classical taxonomic views of *Bacteria* and *Archaea* and phylogenetic models based on the 16S rRNA gene or other universally applicable molecular signals. PCA is a highly robust and efficient unsupervised method of data analysis that could be applied to very large sequence datasets (>10,000) and readily allow visualization of phylogenetic data in a manner that has decided advantages over classical treeing methods. Our early studies revealed that the dimensionality of such data could be reduced to two to three dimensions without a significant loss of information, allowing us to gain insight into the phylogenetic topology defined by the 16S rRNA gene. We also discovered that there were numerous anomalies between the classical taxonomic view of *Bacteria* and *Archaea* and the phylogenetic view that were largely attributable to unresolved synonymies which appeared as outliers in 2D and 3D projections. While useful, PCA was found to have inherent limitations; notably a natural weighting that was attributable to larger taxonomic groups, a variable degree of distortion and rotation that is attributable to the manner in which the reduced dimensions are calculated, and an inability to determine the correct placement of outliers, especially when those outliers were members of minor taxonomic groups.

Subsequently, we discovered that evolutionary distance matrices could be readily viewed as heatmaps within the S+ and R statistical computing environments. This approach had decided advantages over PCA as it allowed us to view the complete data matrices in a distortion-free manner, to visualize different taxonomic arrangements of the data, to pinpoint and correct nomenclatural errors, and to determine what actions were needed to bring the classical taxonomic and phylogenetic views into closer agreement. This phase of the project led to the development of a self-organizing self-correcting classification (SOSCC) algorithm that could pinpoint such anomalies in a semi-automated manner and then optimize the underlying matrix to resolve these anomalies. A prototype was developed in S+ based on the SOSCC algorithm and applied to solve a number of taxonomic problems that have accumulated in the burgeoning 16S rRNA data set, as well as to provide insight into the requirements for deploying this technique as a web service through the RDP.

As the project evolved, it became obvious that neither S+ nor R were sufficiently stable environments for building a client application. To that end, we have re-implemented the SOSCC algorithm in java as an xfire service, optimized the algorithm to provide a more satisfactory user experience (e.g. 30 seconds to produce a maximally smoothed matrix of 1000 sequences), and gained insight into a poorly understood limitation of previous versions of the SOSCC, in which correct placement of some sequences could not be achieved when the algorithm was run in a fully unsupervised, automated version. Work continues on re-implementing the presumptive identification and automatic renaming features of the original algorithm, which will be used to update the next release of the Taxonomic Outline of Bacteria and Archaea and will link the heatmap visualizations to NamesforLife information objects. Deployment of a beta-version SOSCC as an RDP service is planned for the first quarter of 2008. Once deployed, this service will provide valuable information on disagreements between classifications and phylogenies of the prokaryotes and how these problems might be resolved. Furthermore, this experience will provide insights as to how the methodology might be applied to prokaryotes and eukaryotes using other molecular sequences (including complete genomes). The SOSCC may also prove useful in providing insight into expression profiles by pinpointing similarities or discontinuities in microarray data, displayed as optimally smoothed matrices.

* Presenting author

142

GTL

NamesforLife Resolution Services for the Life Sciences

George M. Garrity[1,2]* (garrity@msu.edu), Catherine M. Lyons,[2] and James R. Cole[1,2]

[1]NamesforLife, LLC, East Lansing, Michigan and
[2]Michigan State University, East Lansing, Michigan

Project Goals: NamesforLife (N4L) is an information technology that persistently resolves ambiguity in terminology through the use of a proprietary data architecture that is coupled with persistent identifiers (in the current implementation Digital Object Identifiers are used) and expertly managed terminologies. N4L technology provides transparent links to the occurrence of a technical term or biological name in third party databases or electronic content to managed information about the origins of the term, formal definition, current usage, and related goods and services.

Within the Genomes-to-Life Roadmap, the DOE states that a significant barrier to effective communication in the life sciences is a lack of standardized semantics that accurately describe data objects and persistently express knowledge change over time. As research methods and biological concepts evolve, certainty about correct interpretation of prior data and published results decreases because both become overloaded with synonymous and polysemous terms. Ambiguity in rapidly evolving terminology is a common and chronic problem in science and technology.

NamesforLife (N4L) is a novel technology designed to solve this problem. The core of the technology is an ontology, an XML schema, and an expertly managed vocabulary coupled with Digital Object Identifiers (DOIs) to form a transparent semantic resolution service that disambiguates terminologies, makes them actionable, and presents them to end-users in the correct temporal context. In the first instance, N4L technology has been applied to biological nomenclature, specifically the validly published names of Bacteria and Archaea. These names play a significant role in science, medicine, and government, carry specific meanings to end-users in each of those communities, and can trigger responses that may or may not be appropriate. Biological names also serve as key terms used to index and access information in databases and the scientific, technical, medical, and regulatory literature. Clear understanding of the correct meaning of a biological name, in the appropriate context, is essential. This is a nontrivial task, and the number of individuals with expertise in biological nomenclature is limited. This knowledge can, however, be accurately modeled and delivered through a networked semantic resolution service. Such a service could provide end-users of biological nomenclatures or other dynamic terminologies with the appropriate information, in the correct context, on demand. The same service could also be used by database owners, publishers, or other information providers to semantically enable their offerings, making them discoverable, even when the definition of a name or term has changed.

As proof of principle, a working model of the N4L technology has been built. It allowed us to validate our concepts and gain new insights into previously unaddressed complexities of dynamic vocabularies. The working model also allowed us to introduce the technology to businesses that rely on the proper use of biological names in their product offerings, including scientific publishers, instrument vendors, and other suppliers of information and biological materials. This provided us the opportunity to explore how N4L technology could be applied in various commercial settings to fulfill unmet business needs of vendors and their customers and to do so in a self-supporting manner. The latter goal is achievable because the N4L data architecture is generalizable. The problem that biologists face with terminology, whether it relates to an organism, a gene, or a gene product, is not unique. Analogous problems exist in many other fields.

In this project, we have reduced the working model to a service that can automatically annotate occurrences of names in the scientific literature and databases. The initial target is the *International Journal of Systematic and Evolutionary Microbiology*, the publication of record for all nomenclatural changes for Bacteria and Archaea. To accomplish this objective, we have had to address several technical including transfer of the current model into a more suitable environment to simplify updating and on-the-fly generation of N4L information objects; development of tagging rules to embed links to N4L information objects into on-line content; enabling multiple resolution through the Handle server; development of mini-monographs as an improved human interface to N4L; and development of additional infrastructure to support on-the-fly translation of N4L tagged data in published content.

* Presenting author

143

GTL

The Ribosomal Database Project

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

Center for Microbial Ecology, Michigan State University, East Lansing, Michigan

Project Goals: The Ribosomal Database Project II (RDP) offers aligned and annotated rRNA sequence data and analysis service to the research community. These services help researchers with the discovery and characterization of microbes important to bioenergy production, biogeochemical cycles, and bioremediation.

Through its website (http://rdp.cme.msu.edu/), The Ribosomal Database Project II (RDP) offers aligned and annotated rRNA sequence data and analysis service to the research community (Cole et al., 2007). These services help researchers with the discovery and characterization of microbes important to bioenergy production, biogeochemical cycles, and bioremediation.

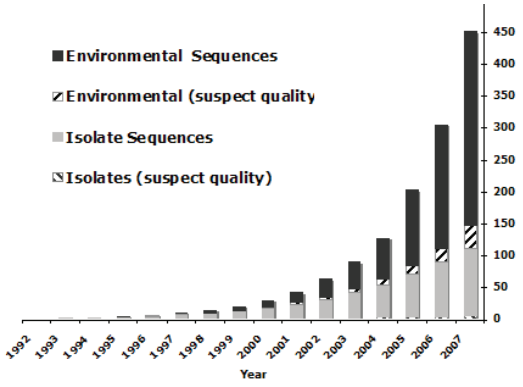


Figure 1. Increase in number of publicly available bacterial small-subunit rRNA sequences. Suspect quality sequences were flagged as anomalous by Pintail in testing with two or more reference sequences from different publications.

Updated monthly, the RDP maintains 451,545 aligned and annotated quality-controlled rRNA sequences as of November 2007 (Release 9.56; Fig. 1). All sequences are tested for sequence anomalies, including chimeras, using Pintail from the Cardiff Bioinformatics Toolkit (Ashelford et al., 2005. *Appl. Environ. Microbiol.* 71:7724-7736). The *myRDP* features introduced last year have grown to support a total of over 2400 active researchers using their *myRDP* accounts to analyze over 460,000 pre-publication sequences in 12,149 sequence library groups.

New Genome Browser: The new RDP Genome Browser allows users to browse information, including rRNA sequences, from DOE and other bacterial genome projects. The Genome Browser includes information about the organism, whether the sequenced strain is recognized as the type strain for the species, and additional links and information provided by the Genomic Standards Consortium (http://darwin.nox.ac.uk/gsc/).

New Taxomatic Visualization Tool: This interactive tool, developed under a separate GTL grant, allows researchers to choose subsets of RDP data and view a distance-based "heatmap" comparison of the user-chosen sequences. With this tool, up to two thousand sequences can be compared at one time. The researcher can interactively zoom-out on the map to gain an overview of the entire data set, or zoom-in to examine specific regions. In addition, taxonomic boundaries can be interactively displayed on the heatmap by manipulating a hierarchy of taxonomic information or by "mousing over" corresponding regions of the heatmap.

Updated Taxonomy: The RDP taxonomy has been updated to reflect changes in release 7.7 of The Taxonomic Outline of Bacteria and Archea (Garrity et al., 2007. *The Taxonomic Outline of Bacteria and Archaea. TOBA Release 7.7, March 2007.* [http://www.taxonomicoutline.org/]).

Expanded Video Tutorials: We have expanded the number of short video tutorials that demonstrate some of the more complex analytical tasks including use of *myRDP*. These tutorials average three minutes in length. They capture the screen as the tasks are performed, while the narrator explains the tasks and the choices available to the user. All tutorials are now available in Flash (with closed-captioning), Quicktime, and Windows media formats.

***Coming Soon*, RDP High-Throughput Pyrosequencing Analysis Pipeline:** New sequencing technologies, such as 454 pyrosequencing, generate tens to hundreds of thousand of partial rRNA sequences at one time and at a much lower per-sequence cost than traditional Sanger sequencing. Sequencing rRNA and other marker genes in environmental samples is a standard method of determining the bacterial composition in the sample. Conventional sequencing technologies are normally too expensive to routinely produce enough sequences to be assured of seeing any but the most abundant sequences. New sequencing technologies such as pyrosequencing have solved this problem and can produce up to hundreds

* Presenting author

of thousands of marker gene sequences from a single sample, enough to provide in-depth of analysis of bacterial composition. However most molecular ecology tools are not able to handle such large numbers of sequences. The RDP is building a Pyrosequencing Pipeline to automate the processing of these large data-sets and provide researchers with the most common ecological metrics, along with the ability to download the processed data in formats suitable for common ecological and statistical packages (Fig 2). As mentioned last year, the RDP Classifier is capable of accurately assigning such short sequences to the bacterial taxonomy (Wang et. al. 2007). In addition, we are developing new tools to align, cluster, dereplicate and simplify the compute-intensive analysis of such large sequencing libraries.

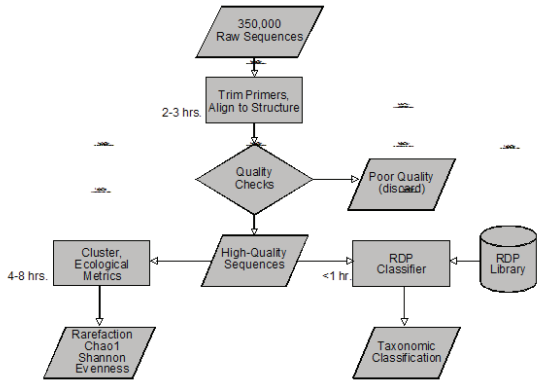


Figure 2. Flowchart showing stages in the RDP High-Throughput Pyrosequencing Analysis Pipeline being developed.

References

1. Cole, J. R., B. Chai, R. J. Farris, Q. Wang, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, A. M. Bandela, E. Cardenas, G. M. Garrity, and J. M. Tiedje. 2007. The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res.* 35 (Database issue): D169-D172; doi: 10.1093/nar/gkl889.
2. Wang, Q, G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl Environ Microbiol.* 73:5261-5267.

144

GTL

The MetaCyc and BioCyc Pathway Databases, and the Pathway Tools Software

Ron Caspi,[1] Carol Fulcher,[1] Pallavi Kaipa,[1] Markus Krummenacker,[1] Suzanne Paley,[1] Lukas Mueller,[2] Anuradha Pujar,[2] Peifen Zhang,[3] Sue Rhee,[3] and **Peter D. Karp**[1]* (pkarp@ai.sri.com)

[1]Bioinformatics Research Group, SRI International, Menlo Park, California; [2]Department of Plant Breeding and Genetics, Cornell University, Ithaca, New York; and [3]Department of Plant Biology, Carnegie Institution, Stanford, California

Project Goals: Develop the MetaCyc database describing experimentally elucidated metabolic pathways and enzymes from many organisms. Develop the BioCyc database describing metabolic pathways predicted from the complete genomes of hundreds of organisms.

Metabolic engineering demands an accurate model of the metabolic network of a target organism and the relationship of that network to the genome, plus powerful analysis tools for constructing, refining, and analyzing that model.

The MetaCyc multiorganism pathway database [1,2] describes experimentally elucidated metabolic pathways and enzymes reported in the experimental literature. MetaCyc is both an online reference source on metabolic pathways and enzymes for metabolic design, and a solid foundation of experimentally proven pathways for use in computational pathway prediction. MetaCyc version 11.6 describes 1,010 pathways from more than 1,000 organisms. The 6,500 biochemical reactions in MetaCyc reference 6,600 chemical substrates, most of which contain chemical structure information. MetaCyc describes the properties of 4,500 enzymes, such as their subunit structure, cofactors, activators, inhibitors, and in some cases their kinetic parameters. The information in MetaCyc was obtained from 15,000 research articles, and emphasizes pathways and enzymes from microbes and plants, although it also contains animal pathways.

Pathway Tools [3,4] constructs a metabolic model of an organism from its annotated genome using the following computational inference tools. The model is in the form of a Pathway/Genome Database (PGDB).

1. It predicts the metabolic pathways of the organism by recognizing known pathways from the MetaCyc database.

* Presenting author

2. It predicts which genes fill holes in those metabolic pathways (pathway holes are pathway steps for which no enzyme has been identified in the genome)
3. It predicts operons for prokaryotic genomes
4. It infers the presence of transport reactions from the names of transport proteins in the genome annotation
5. The software automatically generates a one-screen cellular overview diagram containing the metabolic and transport networks of the cell

A set of graphical editors within Pathway Tools allows scientists to refine a PGDB by adding, or modifying metabolic pathways, gene annotations, reactions, substrates, and regulatory information. The existence of an accurate knowledge base of the metabolic network is a critical resource for metabolic engineering.

The software provides a large number of operations for querying, visualization, web publishing, and analysis of PGDBs. A metabolite tracing tool supports graphical exploration of the path that a substrate follows through the metabolic network, in either the forward or backward direction. The user interactively guides the software in selecting which branches of metabolism to follow, and metabolic paths are highlighted on the cellular overview diagram.

Pathway Tools has a new emphasis on cellular regulation. It can encode information about regulation of transcription initiation and attenuation. Support for more regulatory mechanisms is under development. The Pathway Tools Regulatory Overview can display the entire transcriptional regulatory network of an organism, and can paint omics datasets onto that network to aid in their interpretation.

Pathway Tools has improved support for representing and displaying electron transport events.

Other visualization tools include automated display of metabolic pathways, reactions, enzymes, genes, and operons, and a genome browser.

SRI has applied Pathway Tools and MetaCyc to predict the pathway complements of more than 370 organisms from their complete genomes [1]. The resulting PGDBs are available through the BioCyc.org Web site. In addition, more than 75 groups outside SRI are using Pathway Tools and MetaCyc to produce PGDBs for more than 150 organisms, including the major model organisms for biomedical research (yeast, worm, fly, Dictyostelium), pathogens of biodefense interest, GTL organisms, many other bacteria and archaea, and plants (including Arabidopsis, Medicago, Rice, Tomato, and Potato).

References

1. R. Caspi et al, "The MetaCyc Database of Metabolic Pathways and Enzymes and the BioCyc Collection of Pathway/Genome Databases," *Nucleic Acids Research* in press, 2008 Database Issue.
2. P. Zhang et al, "MetaCyc and AraCyc. Metabolic Pathway Databases for Plant Research," *Plant Physiol* 138:27-37 2005.
3. P.D. Karp et al, "The Pathway Tools Software," *Bioinformatics* 18:S225-32 2002.
4. Paley, S.M. et al, "The Pathway Tools Cellular Overview Diagram and Omics Viewer," *Nucleic Acids Research* 34:3771-8, 2006.
5. Paley, S.M. et al, "Creating Fungal Pathway/Genome Databases Using Pathway Tools," *Applied Mycology and Biotechnology* 6:209-26 2006.

145

GTL

Global Credibility of Sequence Alignments

Bobbie-Jo M. Webb-Robertson[1] (bj@pnl.gov), **Lee Ann McCue**[1]* (leeann.mccue@pnl.gov), and **Charles E. Lawrence**[2] (Charles_Lawrence@brown.edu)

[1]Computational Biology and Bioinformatics, Pacific Northwest National Laboratory, Richland, Washington and [2]Division of Applied Mathematics and the Center for Computational Molecular Biology, Brown University, Providence, Rhode Island

Project Goals: The transcription regulatory network is arguably the most important foundation of cellular function, since it exerts the most fundamental control over the abundance of virtually all of a cell's functional macromolecules. The two major components of a prokaryotic cell's transcription regulation network are the transcription factors (TFs) and the transcription factor binding sites (TFBS); these components are connected by the binding of TFs to their cognate TFBS under appropriate environmental conditions. Comparative genomics has proven to be a powerful bioinformatics method with which to study transcription regulation on a genome-wide level. We will further extend comparative genomics technologies that we have introduced over the last several years, developing and applying statistical approaches to analysis of correlated sequence data (i.e. sequences from closely related

* Presenting author

species). We also plan to combine functional genomic and proteomic data with sequence data from multiple species; combining these complementary data types promises to improve our ability to predict regulatory sites of small or genus-specific regulons.

Genome sequencing initiatives have provided a wealth of bacterial sequence data, including the complete genome sequences of many closely related species like the *Shewanellas*. This rich source of data provides opportunities in comparative genomics to draw inferences on species phylogeny, metabolic capabilities, and regulation; however it also presents challenges associated with high-dimension solution space. For instance, sequence alignment is fundamental to the analysis of genome data, yet sequence alignment methods commonly focus on identifying only the best possible alignment between two sequences. However, when a single alignment is chosen for the comparison of two (or more) sequences, it is a point estimate selected from a large ensemble of all possible alignments. For example, two small sequences of length 20 generate over 2.7e+29 possible local alignments. Given the immense size of the alignment space, it is not surprising that the most probable alignments, and thus all individual alignments, often have very small probabilities.

This finding raises three questions:

1. In discrete spaces, how strongly does the available data recommend a single chosen alignment?
2. When the data provide weak evidence for any single alignment, what criteria can be used to judge the credibility, and what are reasonable limits in the degree of variation within the ensemble of alignments, that are consistent with the data?
3. How can we identify the single alignment that best represents the ensemble of alignments, and that is consistent with the data?

We present results in support of the following answers to these questions:

1. The strength of the recommendation of the data for any specific alignment is equal to its posterior probability under the assumed probabilistic model.
2. A credibility limit is the radius of the smallest hypersphere around a proposed alignment that contains a specified proportion of the posterior weighted ensemble, where the radius is measured by the number of elements by which two solutions differ. The size of this limit characterizes an alignment's credibility.
3. The alignment with the minimum credibility limit best represents the ensemble.

We find high variability in the credibility limits in the alignments of promoter sequences from closely related species. In addition, we find that the alignment with the minimum credibility limit (the ensemble centroid alignment) often differs significantly from a single "best" alignment (maximum similarity alignment). Furthermore, often the credibility limit of the maximum similarity alignment is no better than that of an alignment selected at random from the posterior weighted alignment space. We have demonstrated that credibility limits can be used to define criteria for the alignment of orthologous promoter regions from *Shewanella* species prior to motif prediction.

* Presenting author

Communication

146

GTL

Genome Management Information System: A Multifaceted Approach to DOE Genome Research Facilitation and Communication

Anne E. Adamson, Shirley H. Andrews, Jennifer L. Bownas, Kris Christen, Sheryl A. Martin, Marissa D. Mills, Kimberly McGrew, Suzanne Parete-Koon, Judy M. Wyrick, Anita J. Alton, and **Betty K. Mansfield*** (mansfieldbk@ornl.gov)

Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee

Project Goals: Help build the critical multidisciplinary community needed to advance systems biology research for DOE energy and environmental missions and foster industrial biotechnology. The Genome Management Information System (GMIS) contributes to DOE Genomics:GTL program strategies and communicates key GTL scientific and technical concepts to the scientific community and the public. We welcome ideas for extending and improving communications and program integration to represent GTL science more comprehensively.

Concerted communication is key to progress in cutting-edge science and public accountability. With support from the Department of Energy's (DOE) Office of Science, the Genome Management Information System (GMIS) has for 19 years been the main communication resource supporting the pioneering Human Genome Project (HGP). However, since 2000 our primary focus has shifted to helping plan, implement, and communicate DOE's Genomics:GTL (GTL, formerly Genomes to Life) program enabled by the HGP. The goal of GTL is to attain a predictive, whole systems–level understanding of microbes and plants to help enable biobased solutions to DOE missions. Our mission is to work with DOE staff and the broad scientific community to communicate biological science challenges and findings to stimulate advances at interdisciplinary interfaces, democratize access to the growing bounty of resources and data, and drive more-informed scientific and societal discourse. Our goals focus on three areas: (1) facilitate GTL planning, research, and communication; (2) respond to communication needs of related projects; and (3) continue to communicate about DOE genomics research and potential applications.

Technical communication integrating all facets of GTL research is critical for spurring innovation at the most rapid pace and at the lowest cost. Such communication is important to achieving DOE missions and, ultimately, fostering U.S. competitiveness through growth in the industrial and environmental sector of the estimated trillion-dollar biotechnology industry spurred largely by DOE genomics research.

Throughout the HGP (1989 to 2003), GMIS strategic networking and communication helped promote collaborations and contributions from numerous fields and reduced duplicative scientific work in the growing genomics community. GMIS staff and the resources we created became the primary "go-to" source for information on all things genomic for much of the scientific world, the media, and the public. A large collection of award-winning, informative literature; websites; large-format exhibits; and graphics forms the core of these resources, which are assessed frequently for value, timeliness, and cost-effectiveness. Hundreds of thousands of document copies have been distributed. In addition, last year alone, GMIS websites received some 20 million page views (224 million hits), many from people who are just learning of the HGP through new textbooks or news coverage of the latest gene discovery. Through our resources, networking at various professional scientific and related education meetings, and strategic partnerships, we continue to broaden our reach and focus the attention of numerous people in the national media, government, academia, industry, education, and medicine on DOE genomics programs.

For the scientific community, communication and research information integration are even more important for GTL than for the HGP, which relied on one dominant technology—DNA sequencing—and produced one major data set—DNA sequence. This next generation of biology is far more complex and involves a wider array of technologies, many just emerging, with new types of data sets that must be available to a larger, more diverse research community. Moreover, disparate groups of interdisciplinary scientists must be engaged to achieve the productive dialogue leading to research endpoints that will ensure the success of GTL. The stakes are high: GTL resources and data have the potential to enlarge

* Presenting author

the research community working on biotechnological approaches to DOE missions, resulting in more rapidly evolving scientific thinking and progress in these and related areas of critical global importance. The extensive experience, scientific knowledge, and professional credibility gained during the HGP years now place GMIS in a prime position helping tackle an even greater challenge: Helping the Office of Biological and Environmental Research—within the Office of Science—develop the scientific understanding, data, and tools for GTL systems biology in support of DOE mission applications. Communication strategies must be dynamic and evolve along with programmatic needs.

Since 2000, GMIS GTL communication and research integration strategies have included helping facilitate numerous scientific workshops to develop GTL program plans; producing GTL symposia at national scientific meetings; and creating numerous informational resources and tools used by scientists, program administrators, and others. Two important research plans we have produced with the research community are: *DOE Genomics:GTL Roadmap: Systems Biology for Energy and Environment* (August 2005) and *Breaking the Biological Barriers to Cellulosic Ethanol: A Joint Research Agenda* (June 2006). Other work in progress this fiscal year includes a new GTL strategic plan, an overview of the DOE Bioenergy Research Centers, a climate plan and a plan emanating from a carbon cycle workshop. We also plan to enhance the GTL web presence.

In addition to helping drive communication within the scientific community, GMIS will continue to leverage the high level of public interest in the HGP and genomic science with our established and future resources to inspire a similar wonder at the challenging new task before us: Learning how genomic "parts" (i.e., genes and regulatory components) work together to produce the processes of life. GTL pursues this grand scientific challenge via investigations in microbial and plant systems, whose sophisticated biochemical abilities have yet to be understood and tapped. We will communicate the excitement of these investigations and their potential applications beyond the interested research community to broader audiences.

> "Interdisciplinary research…is a mode of research by teams or individuals that integrates information, data, techniques, tools, perspectives, concepts, and/or theories from two or more disciplines or bodies of specialized knowledge to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single discipline or area of research practice." [National Academy of Sciences, National Academy of Engineering, and Institute of Medicine of the National Academies, *Facilitating Interdisciplinary Research*, The National Academies Press, Washington, D.C., 2005.]

* Presenting author

Ethical, Legal, and Societal Issues

147

GTL

Meeting the Demand for Biofuels: Impact on Land Use and Carbon Mitigation

Madhu Khanna (khanna1@uiuc.edu), Atul Jain,* Hayri Onal, and Jurgen Scheffran

University of Illinois, Urbana-Champaign, Illinois

Project Goals: We propose to conduct an integrated, interdisciplinary research project that will investigate the implications of large scale production of biofuels for land use, crop production, farm income and greenhouse gases. In particular, this research will examine the mix of feedstocks that would be viable for biofuel production and the spatial allocation of land required for producing these feedstocks at various gasoline and carbon emission prices as well as biofuel subsidy levels. The implication of interactions between energy policy that seeks energy independence from foreign oil and climate policy that seeks to mitigate greenhouse gas emissions for the optimal mix of biofuels and land use will also be investigated. The objectives of this research are to (a) determine yield and greenhouse gas mitigation benefits, in the form of soil carbon sequestration and displacement of carbon emissions from gasoline, of each type of feedstock, (b) examine the optimal allocation of existing cropland for feedstock production, the mix of feedstocks that should be produced and the spatial pattern of land use in the U.S. at various expected prices of gasoline, market prices of carbon emissions, and biofuel subsidy levels, and (c) investigate the optimal plant sizes, transportation patterns and areas to locate bio-refineries.

We propose to conduct an integrated, interdisciplinary research project that will investigate the implications of large scale production of biofuels for land use, crop production, farm income and greenhouse gases. In particular, this research will examine the mix of feedstocks that would be viable for biofuel production and the spatial allocation of land required for producing these feedstocks at various gasoline and carbon emission prices as well as biofuel subsidy levels. The implication of interactions between energy policy that seeks energy independence from foreign oil and climate policy that seeks to mitigate greenhouse gas emissions for the optimal mix of biofuels and land use will also be investigated. This project will thereby contribute to the ELSI research goals of sustainable biofuel production while balancing competing demands for land and developing policy approaches needed to support biofuel production in a cost-effective and environmentally friendly manner.

The objectives of this research are to (a) determine yield and greenhouse gas mitigation benefits, in the form of soil carbon sequestration and displacement of carbon emissions from gasoline, of each type of feedstock, (b) examine the optimal allocation of existing cropland for feedstock production, the mix of feedstocks that should be produced and the spatial pattern of land use in the U.S. at various expected prices of gasoline, market prices of carbon emissions, and biofuel subsidy levels, and (c) investigate the optimal plant sizes, transportation patterns and areas to locate bio-refineries.

We will undertake this research by integrating a biophysical model, Integrated Science Assessment Model (ISAM), and an economic model, Agricultural Policy Analysis Model (APAM), together with detailed GIS data on soil and climate at a 1km x 1km level for all corn producing states in the U.S. ISAM will be used to simulate yields of various row crops, such as corn, soybeans and wheat, as well as two bioenergy crops, switchgrass and miscanthus. It will also be used to determine the carbon sequestration rates with alternative land uses and tillage practices and the life-cycle carbon emissions of alternative feedstocks. Agro-zones with homogeneous growing conditions will be defined using the crop productivity estimates obtained from ISAM. Profitability of alternative land uses and tillage management choices in each zone will be determined. These together with the zone specific information on life cycle carbon emissions of alternative feedstocks will be used to determine the profit maximizing land use decisions for each agro-zone. APAM embeds a dynamic regional land use allocation model within a national model with demand and supply for row crops that determines equilibrium prices. The regional and the national models are solved iteratively each time period, with equilibrium outcomes in a given year affecting land allocation decisions in the next. These will be used to determine optimal land allocation decisions over a 20-year horizon. Following the determination of the optimal land use allocation for feedstock production, we will develop a location choice model that will identify the optimal locations for bio-refineries and optimal transportation pattern. Based on the distances (thus, transportation costs) from production sources

* Presenting author

of the feedstock and the railroad/road network, certain locations may be more suitable for corn and corn-stover based biofuel plants, while others may be more suitable for producing biofuels by using perennial grasses. The framework developed here will be used to examine the policy implications of various levels of biofuel subsidies, carbon prices and gasoline prices for land allocation between food and fuel production as well as for greenhouse gas emissions.

148

GTL

The Biofuels Revolution: Understanding the Social, Cultural, and Economic Impacts of Biofuels Development on Rural Communities

Theresa Selfa[1]* (tselfa@ksu.edu), Richard Goe,[1] Laszlo Kulcsar,[1] Gerad Middendorf,[1] and Carmen Bain[2]

[1]Department of Sociology, Kansas State University, Manhattan, Kansas and [2]Department of Sociology, Iowa State University, Ames, Iowa

Project Goals: The goal of this project is to provide a better understanding of the socio-economic and cultural implications of biofuels development for rural communities, and to contribute to more informed policy development regarding bioenergy. This goal will be accomplished through an in-depth analysis of the impacts of ethanol production on six rural communities in the Midwestern states of Kansas and Iowa.

A new wave of economic growth is currently sweeping across rural communities in the Midwest region of the U.S., fueled by the construction and expansion of ethanol biorefineries and the expansion of biofuel crop production. While the expansion of the biofuels industry promises to bring jobs and economic vitality to rural communities, it is also creating dilemmas for farmers and rural communities in weighing the benefits of income growth and job growth against safety risks, increased pollution, and the potential of overextending water supplies. Presently, there is little empirical knowledge about the social, cultural and economic impacts of biofuels development on rural communities. This research is intended to help fill these lacunae through an in-depth analysis of the social, cultural, and economic impacts of ethanol biorefinery industry on six rural communities in the Midwestern states of Kansas and Iowa. The goal of this project is to provide a better understanding of the socio-economic and cultural implications of biofuels development for rural communities, and to contribute to more informed policy development regarding bioenergy.

Research Questions:

1. To understand how the growth of biofuel production has affected and will affect Midwestern farmers and rural communities in terms of economic, demographic, and socio-cultural impacts.
2. To determine how state agencies, groundwater management districts, local governments and policy makers evaluate or manage bioenergy development in relation to competing demands for economic growth, diminishing water resources, and social considerations.
3. To determine the factors that influence the water management practices of agricultural producers in Kansas and Iowa (e.g. geographic setting, water management institutions, competing water-use demands as well as producers' attitudes, beliefs, and values) and how these influences relate to bioenergy feedstock production and biofuel processing.
4. To determine the relative importance of social-cultural, environmental and/or economic factors in the promotion of biofuels development and expansion in rural communities.

Research Methodology

The comprehensive methodology will include: demographic analysis; field research involving in-depth personal interviews and focus groups with key informants and selected community groups; a general population opinion survey of community residents; and a content analysis of local newspapers and print media. These four methodological procedures will generate an extensive and detailed database of the nature of biorefinery development in the six rural communities serving as case study sites. Where possible, the findings derived from one method will be "triangulated" with the findings derived from other methods. Taken together, the data resulting from this study will provide a more detailed and comprehensive understanding of the social, economic, and cultural impacts of ethanol biorefineries on rural than has heretofore been developed. The research will lead to completion of six detailed case studies of rural communities that are current or planned locations for ethanol biorefineries.

Preliminary Findings

A substantial number of rural communities in Kansas and Iowa have undergone decades of stagnation and decline in terms of the size of their population and

* Presenting author

employment base. A key policy issue is whether ethanol refineries are being located in those rural communities where they could have the most beneficial impact in terms of attracting population, generating new jobs and increasing incomes? Preliminary findings indicate that as of October, 2007, ethanol refineries have been located in non-metropolitan counties that had significantly lower poverty rates in 2000 compared to those rural counties without a refinery. In addition, non-metropolitan counties where ethanol refineries have located were found to have: (a) significantly smaller percentages of households that are working poor; (b) significantly greater levels of aggregate income; (c) significantly greater increases in real aggregate income over the 1989-1999 period; (d) significantly higher median household incomes; (e) significantly larger employment bases; (f) significantly larger total populations; (g) significantly more urbanized populations; and (h) significantly higher percentages of crop acreage devoted to corn production. These findings suggest that at this juncture, ethanol refineries are serving to reinforce the structural advantage of larger, more urbanized non-metropolitan communities with greater wealth and larger economies. In turn, this implies that unless this pattern changes, smaller non-metropolitan communities within the two state area will be at a further disadvantage as a result of biofuels development.

For Additional Information:

Project information and research findings will be available at: http://www.ksu.edu/sasw/kpc/biofuels/project_doe.htm.

149

GTL

Analysis of Global Economic and Environmental Impacts of a Substantial Increase in Bioenergy Production

Wallace E. Tyner (wtyner@purdue.edu), Thomas W. Hertel, Farzad Taheripour,* and Dileep K. Birur

Department of Agricultural Economics, Purdue University, West Lafayette, Indiana

Project Goals: The goal of this research is to develop realistic assessments of the economic and environmental impacts of regional and global policies designed to stimulate bioenergy production and use. We will build on the unique strengths of the Global Trade Analysis Project (GTAP) to analyze economic impacts of alternative bioenergy policies at regional and global levels. We will use the Terrestrial Ecosystems Model (TEM) model to evaluate the potential for new lands to be brought into production in the wake of biofuel programs, as well as to validate environmental consequences of these policies and check their feasibility from a fundamental bio-geochemical perspective.

Introduction

With the growing concerns on energy security and climate change, biofuels have witnessed rapid increase in production in rich countries particularly in the United States and the European Union. Massive subsidies offered by these countries, have resulted in large-scale implementation of biofuels programs which have profound global economic, environmental, and social consequences. Current studies do not provide much insight into how alternative bioenergy production scenarios could change global agricultural production nor the incomes of affected groups both within and across nations. The increasing importance of biofuels and lack of comprehensive studies on global impacts have opened up several research avenues. Since biofuels are produced mainly from agricultural sources, their effect is largely felt in agricultural markets and land-use, with repercussions for international trade. As the World Bank reports, nearly 70 percent of the world's poor live in rural areas in developing countries and derive their primary livelihood from agriculture. Higher biofuel feedstock prices can help the farmers who grow them, but an increase in food prices can hurt the poor as they spend a large share of their budget on food. Keeping these issues in view, this project aims to develop a realistic assessment of the economic and environmental impacts of regional and global policies designed to stimulate biofuels production and use. This project will be completed over a three-year period. The project builds on the unique strengths of the Global Trade Analysis Project (GTAP) based at Purdue University.

Analytical Framework

A Computable General Equilibrium (CGE) modeling in GTAP framework (Hertel, 1997) is adopted as it is best suited for studying the global, socio-economic impacts of biofuel technologies and policies. The GTAP database and its analytical framework are widely used for analysis of global trade, energy, environment, technical change, and poverty issues in an economy-wide context. In order to evaluate alternative bioenergy scenario impact on land-use, potential land for feedstock and other crops production, water availability and greenhouse gas (GHG) emissions, we utilize the Terrestrial Ecosystem Model (TEM), a widely used model in ecological research. The

* Presenting author

resulting factor earnings and commodity prices from the interaction of economic and environment modules are translated through the poverty module to determine the change in poverty headcount, by population strata in a sample of 15 developing countries. In all, through this project, we will be able to empirically evaluate the impacts of biofuel policies in the U.S., EU, and Brazil on the global economy, focusing on impacts on agricultural markets, global poverty and environment.

Research Progress and Preliminary Results

We have incorporated three explicit biofuels sectors such as corn-based ethanol, sugarcane-based ethanol, and vegetable oil based biodiesel, in to the GTAP data base (Dimaranan, 2007 and Taheripour et al., 2007). In order to introduce biofuels as energy substitutes/complements into the GTAP model, we use GTAP-Energy model (Burniaux and Truong 2002; McDougall and Golub, 2007) linking with Agro-ecological Zones (AEZs) (Lee et al., 2005) for each of the land using sectors. The GTAP-E model with biofuels and AEZs provides a clear picture regarding the impacts of growing importance of biofuels on global changes in crop production, utilization, prices, factor movements, trade, land-use change etc. For validation of the model, we project a hypothetical biofuel economy forward in time and compare the model predictions with historical evidence from 2001 to 2006. We focus on three main drivers of biofuel boom in the U.S.: hike in crude oil prices, replacement of MTBE by ethanol in gasoline additives, and subsidy for ethanol. Using this historical simulation, we calibrate the key elasticity of energy substitution between biofuels and petroleum products in each region. With these parameter settings, the model does a reasonably good job of predicting the share of feedstock in biofuels and related sectors in the major biofuel producing regions.

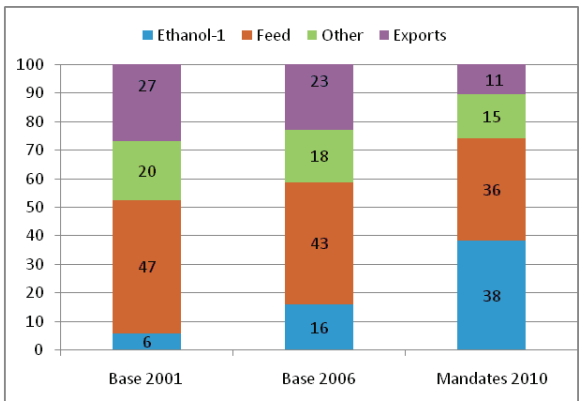


Figure. Disposition of Coarse grains in the U.S. (%)

As an illustration, we analyze the impact of implementation of biofuel mandates for 2010 in the U.S., and EU. The European Union has set the goal to reach a biofuel share of 5.75% in transportation fuels market by 2010. The United States is considering a target to reach about 35 billion gallons of biofuels by 2017. We adopt these targets for a "mandates" simulation for the year 2010. Both of the regions are expected to substantially increase the share of agricultural products utilized by the biofuels sector. For example, the corn share in ethanol production could double from 2006 levels in the U.S. (Figure). Similarly, share of oilseeds going to biodiesel in the EU triple with doubling the price of these feedstocks from 2006 levels and sharply reducing their exports from these two regions. In the EU, the majority of the biodiesel driven demand for oilseeds is met from imports, with import volume rising by more than $4 billion. These increases in biofuels demand also have profound influence on agricultural production and land-use. An 11% increase in corn acreage from 2006 in the U.S. results in significant falls in wheat and soybean acreage along with other crops, livestock, and forestry land use. The acreage devoted to oilseeds in the EU increases by 21%, which comes at the cost of forestry and other crops. The combined impact of U.S. and EU biofuel mandates puts considerable pressure on agriculture and forest lands throughout the world. Due to slight decline in import of petroleum products, the trade balance improves for the U.S. Overall, the biofuel mandates in the U.S. and EU are likely to have significant and lasting impacts on the global pattern of agricultural production and trade.

References

1. Burniaux, J. and T. Truong (2002) "GTAP-E: An Energy-Environmental Version of the GTAP Model", GTAP Technical Paper No. 16, Center for Global Trade Analysis. Purdue University, West Lafayette, IN.
2. Dimaranan, B.V., Edt. (2007). Global Trade, Assistance, and Production: The GTAP 6 Data Base, Center for Global Trade Analysis, Purdue University, West Lafayette, IN 47907, USA.
3. Hertel, T. W. (1997) "Global Trade Analysis, Modeling and Applications." Cambridge University Press, Cambridge.
4. Lee, H.-L., T. W. Hertel, B. Sohngen and N. Ramankutty (2005) "Towards and Integrated Land Use Data Base for Assessing the Potential for Greenhouse Gas Mitigation." GTAP Technical Paper # 25, Center for Global Trade Analysis, Purdue University, https://www.gtap.agecon.purdue.edu/resources/res_display.asp?RecordID=1900

* Presenting author

5. McDougall, R. and Golub, A. (2007 forthcoming) "GTAP-E Release 6: A Revised Energy-Environmental Version of the GTAP Model", GTAP Technical Paper, forthcoming.
6. Taheripour, F., D.K. Birur, T.W. Hertel, and W.E. Tyner (2007) 'Introducing Liquid Biofuels into the GTAP Database' GTAP Research Memorandum. Center for Global Trade Analysis, Purdue University, West Lafayette, IN 47907, USA.

150

GTL

Intellectual Property and U.S. Public Investments in Research on Biofuel Technologies

Kerri Clark* (kleclark@ucdavis.edu), Rohan Patel, Kyle Jensen, and **Alan Bennett**

Public Intellectual Property Resource for Agriculture (www.pipra.org), University of California, Davis, California

Project Goals: PIPRA will map the intellectual property landscape of biofuels with a focus on "cellulosic" ethanol. We will determine the different technological approaches to making biofuels, the degree to which these approaches are covered by patents, and show how patent ownership is distributed across the public institutions and industries. Further, we will assess possible strategies, such as research consortia sharing agreements, patent license clearinghouses and patent pools, to co-ordinate access to multiple technology components considered necessary for future developments.

Increased interest in the replacement of fossil fuels with biofuels to combat global warming and increase national security has resulted in a surge in biofuel research whose outcomes are adding to an already complex intellectual property (IP) landscape. An understanding of the biofuel IP landscape can be used to better inform policy makers, sponsors, institutions and researchers to promote and conduct commercially viable research which will support the maximization of returns on research investments. To increase this understanding we, at the Public Intellectual Property Resource for Agriculture (PIPRA) group, are mapping the IP landscape of biofuel technologies focusing on bioethanol production from cellulosic biomass. This landscape will be used to analyze global patenting activity including identifying the predominant patent applicants, technology advances and geographical patenting trends.

The IP landscape has been divided into sections encompassing the bioethanol production stages from the farm (including seeds/germplasm, farming techniques, storage and transport of biomass) to the ethanol plant (comprising the conversion of biomass to fermentation feedstock, fermentation and ethanol recovery). Current efforts are directed to determining the landscape associated with the enzymes involved in the release of fermentable sugars from the cellulosic biomass; a key step where production costs could be substantially reduced thereby increasing the overall energy return on investment. Preliminary analysis has revealed more than 90% of patents and patent applications have been awarded to private companies, reflecting the pattern of private relative to public funding for research in this area until recently. Furthermore, associated with changes in public policy and the increase in biofuel research, patenting activity has increased noticeably since 2002. This information can be used to form an understanding of the overall IP landscape and the need for licensing agreements for enabling technologies as well as assisting in the design of research projects with maximum freedom-to-operate.

PIPRA is a not-for-profit organization whose objective is to support innovation in public sector agriculture research institutes for commercial and humanitarian uses, by providing a wide range of technical services for improved IP management. These services include the provision of enabling technologies, generation and analysis of IP landscapes, educational services and the facilitation of licensing and material transfer agreements with member institutions. PIPRA comprises 45 institutional members in 14 countries.

151

GTL

Integrating ELSI into the Center for Nanophase Materials Sciences at the Oak Ridge National Laboratory

Amy K. Wolfe* (wolfeak@ornl.gov) and **David J. Bjornstad** (bjornstaddj@ornl.gov)

Oak Ridge National Laboratory, Oak Ridge, Tennessee

Project Goals: This project seeks to incorporate an Ethical, Legal, and Societal Implications (ELSI) activity into Oak Ridge National Laboratory's (ORNL's) Center for Nanophase Materials Sciences (CNMS) in a manner that is forward-looking, consistent with the

* Presenting author

conduct of nanoscience and technology research, and valuable to the user communities of the CNMS and the other nanoscience centers.

Despite explicit language in the *21st Century Nanotechnology Research and Development Act of 2003* that encourages the integration of ELSI (Ethical, Legal, and Societal Implications) research and activities with nanotechnology research and development activities, it is not immediately obvious how to do so effectively. By "effectively," we mean adding value but not burden to the conduct of science, taking advantage of proximity to the center to carry out ELSI research, and adopting a neutral position toward the science and its outcomes. We anticipate that an effective integration would enhance the experience obtained by scientists through their participation in a nanotechnology research center, improve the quality of ELSI research, and, potentially, identify new nanotechnology research topics.

This presentation describes a project that seeks to incorporate an ELSI activity into Oak Ridge National Laboratory's (ORNL's) Center for Nanophase Materials Sciences (CNMS) in a manner that is forward-looking, consistent with the conduct of nanoscience and technology research, and valuable to the user communities of the CNMS and the other nanoscience centers. The project responds to the abovementioned *Nanotechnology Act of 2003* and to remarks by a senior DOE official in open forum, both of which encourage the integration of ELSI with nanotechnology research and development activities. To achieve our objectives, we are: (1) conducting societal implications research on selected issues of importance to nanoscience and technology; (2) eliciting and providing ELSI-related information to the CNMS and its users and to the broader nanotechnology community; and (3) developing and implementing an ELSI consultation resource that creates the capacity to apply lessons from the broad ELSI and nanotechnology-ELSI (NELSI) literature to the specific issues and concerns that that could impinge on the conduct of nanotechnology research. We are working closely with key CNMS personnel to assure that ELSI is incorporated effectively.

This presentation describes the structure of our ELSI endeavor, our methods, initial efforts, and anticipated outcomes. It emphasizes our efforts to integrate and communicate with the CNMS and nanoscience/nanotechnology research communities, and begins to identify salient ELSI topics, such as issues associated with the transition from laboratory to commercial settings and intellectual property rights.

152

GTL

Lessons from Experience about Societal Responses to Emerging Technologies Perceived as Involving Risks

Thomas J. Wilbanks[1]* (wilbankstj@ornl.gov), Susan Cozzens,[2] Brian Davison,[1] Paul Gilman,[3] Eugene Rosa,[4] and Paul Stern[5]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [2]Georgia Institute of Technology, Atlanta, Georgia; [3]Oak Ridge Center for Advanced Studies, Oak Ridge, Tennessee; [4]Washington State University, Pullman, Washington; and [5]National Research Council

The objectives of this project are (a) to consider the historical experience in the United States with developing new technologies associated with public concerns about risk, and (b) to consider how lessons learned from this experience might be relevant to societal implications of emerging technologies involving bioengineering for alternative energy production.

The research team will utilize established analytic-deliberative group processes to extract key lessons learned from the experience with "risky" technologies in the U.S. and from inferences from relevant social science literatures, producing a report on lessons learned that should be applicable to a wide range of emerging technologies. Next, it will conduct a prototype test of the possible relevance of these lessons to long-term bioengineering science and technology for innovative bioenergy production. It will then review the results and invite participation by a wider community of experts at a national workshop on the topic of the project, in collaboration with DOE and the Oak Ridge Center for Advance Studies (ORCAS).

* Presenting author

Appendix 2: Websites

Program Websites

- Genomics:GTL: http://genomicsgtl.energy.gov
- This book: http://genomicsgtl.energy.gov/pubs/2008abstracts/
- *Breaking the Biological Barriers to Cellulosic Ethanol: A Joint Research Agenda* http://genomicsgtl.energy.gov/biofuels/b2bworkshop.shtml
- DOE Microbial Genome Program: http://microbialgenomics.energy.gov
- DOE and USDA Biomass Genomics Research http://genomicsgtl.energy.gov/research/DOEUSDA/

Bioenergy Research Center Websites

- Overview http://genomicsgtl.energy.gov/centers/
- BioEnergy Science Center: http://www.bioenergycenter.org
- Great Lakes Bioenergy Research Center: http://www.greatlakesbioenergy.org
- Joint BioEnergy Institute: http://www.jbei.org

Project and Related Websites

- CABIN: http://www.sysbio.org/capabilities/compbio/cabin.stm
- DOE Joint Genome Institute: http://jgi.doe.gov
- Gaggle: http://gaggle.systemsbiology.org/docs/
- MAGGIE: http://masspec.scripps.edu/maggie/
- MicrobesOnline: http://microbesonline.org
- Microbial Protein-Protein Interaction Database: http://MiPPI.ornl.gov
- PNNL Proteomics: http://ober-proteomics.pnl.gov
- Public Intellectual Property Resource for Agriculture (PIPRA) http://www.pipra.org
- RegTransBase: http://regtransbase.lbl.gov
- Ribosomal Database Project II (RDP): http://rdp.cme.msu.edu
- *Shewanella* Knowledgebase: http://shewanella-knowledgebase.org
- Socioeconomic Impacts of the Biofuels Revolution: http://www.ksu.edu/sasw/kpc/biofuels/project_doe.htm
- The SEED: http://theseed.uchicago.edu
- Virtual Institute for Microbial Stress and Survival: http://vimss.lbl.gov

Author Index

Indexed by Page Number

C

D

E

F

Indexed by page number.

G

H

Indexed by page number.

I

J

K

L

Indexed by page number.

M

N

O

P

Indexed by page number.

Q

R

S

Indexed by page number.

T

U

V

W

Indexed by page number.

X

Y

Z

Indexed by page number.

Institutional Index

Indexed by Page Number

Indexed by page number.

The complimentary use of small angle x-ray scattering with crystallography in the determination of biological macromolecular structures.

John Tainer, Greg Hura
Lawrence Berkeley National Laboratory

Project_Goals: Recent advances in small angle x-ray scattering (SAXS) technique and analysis have enabled shape prediction of proteins in solution. The SAXS technique is particularly powerful in combination with partial high resolution structures. SAXS can efficiently reveal the spatial organization of protein domains, including domains missing from or disordered in known crystal structures, and establish cofactor or substrate-induced conformational changes. Following a short introduction to SAXS, examples from data collected at SIBYLS, a dual SAXS and protein crystallography synchrotron beamline, will be drawn upon to demonstrate the complimentary use of SAXS with protein crystallography. I will also describe the recent implementation of a sample loading automation tool for true high throughput SAXS data collection. Several examples of the utility of high throughput SAXS will discussed in the context of the DOE/Gtl funded program project MAGGIE (Molecular Assemblies, Genes, and Genomics Integrated Efficiently).

A core aim of the MAGGIE (Molecular Assemblies, Genes, and Genomics Integrated Efficiently) program project is to develop technologies for proteomic scale visualization of macromolecular structure. While determining function from structure remains a challenge, structure aids in understanding how macromolecules function. Detailed macromolecular structure determination from X-ray crystallography or NMR has provided a broad and deep survey of soluble biomolecules. Several techniques, both experimental and computational, exploit this information to provide significant insight into and prediction of structures which have not been probed by these often challenging atomic resolution techniques. Solution X-ray scattering (also known as SAXS: small angle X-ray scattering) has matured to provide structures at a 10Å resolution. This resolution is sufficient to elucidate a great deal of the architecture of macromolecules, how they interact and exchange products along a pathway. SAXS may be powerfully combined with information from atomic resolution and computational structure prediction methods. Here we detail our highthroughput data collection, data analysis, and data storage programs and strategies. Full SAXS data collection (including collection of buffer blanks and a 3 fold serial dilution) on a macromolecule of interest is complete in under 10 minutes. A computer processor limited data analysis tree has been developed requiring minimal human intervention. Our web accessible data storage utility BioIsis (www.bioisis.net) allows direct access to data at all stages of analysis. We demonstrate this strategy on 16 protein from pyrococus furiosis previously prepared for a crystallographic based structural genomic effort.

Redox balancing pathways in RubisCO knockout mutants of nonsulfur purple photosynthetic bacteria and the potential for enhanced biohydrogen production

Rick A. Laguna*[1] (laguna.2@osu.edu), James C. Liao[2], and **F. Robert Tabita**[1]

[1]Department of Microbiology, The Ohio State University, 484 W. 12th Avenue, Columbus, OH 43210 USA; [2]Department of Chemical Engineering, 5531 Boelter Hall, University of California at Los Angeles, Los Angeles, California 90095

Efficient bacterial metabolism and cellular integrity requires redox homeostasis. This is particularly important in nonsulfur purple (NSP) photosynthetic bacteria that generate reductant via light reactions, as well as via the oxidation of organic carbon compounds. In these organisms, redox balance is maintained via multifaceted and integrated regulatory networks that enable these organisms to respond to diverse and highly variable inter- and intracellular environments. When growing photoheterotrophically, especially with reduced carbon sources and in the absence of exogenous reductant, excess reducing equivalents are normally consumed via the reduction of CO_2. Other processes, such as nitrogen fixation, allow N_2 to compete with CO_2 as the terminal electron acceptor. Over the years (1-4), we have shown that NSP photosynthetic bacteria possess an array of metabolic and regulatory capabilities that allow for the utilization of alternative redox sinks when the primary electron sink, CO_2, is nullified via the inactivation or deletion of the RubisCO genes. RubisCO is the key enzyme of the Calvin-Benson-Bassham pathway of CO_2 assimilation. In order to grow photoheterotrophically, such RubisCO-compromised strains develop interesting strategies and alter their basic metabolic profile. For example, in many instances the derepression of nitrogenase synthesis occurs under normal repressive conditions. Such gain-of-function adaptive mutant strains have been obtained from *Rhodobacter capsulatus, Rhodobacter sphaeroides, Rhodospirillum rubrum*, and *Rhodopseudomonas palustris,* whereby such strains balance their redox potential via nitrogenase-catalyzed reduction of protons to hydrogen gas (1-4).

The use of hydrogen as an energy source is attractive because the end-product is H_2O, as opposed to CO_2, a green house gas generated by the burning of fossil fuels. H_2 is normally produced by nonsulfur purple photosynthetic bacteria during nitrogen limiting conditions by the nitrogenase complex. However, the gain-of-function adaptive mutant strains of NSP bacteria are able to produce copious quantities of H_2. Here we have compared the amount of hydrogen produced from the RubisCO-compromised *R. capsulatus, R. sphaeroides, R. rubrum, and R. palustris* RubisCO mutant strains and have shown that the levels of H_2 produced are significantly greater than previously reported. We are also investigating the integrative control of CO_2 and N_2 fixation, specifically, how normal control of the nitrogenase complex is regained when a functional CBB cycle is restored in such strains.

In order to maximize hydrogen production, JCL is developing a modeling approach to allow us to target certain aspects of metabolism that may be altered via metabolic engineering. Metabolic engineering requires selecting enzymes for expression tuning in order to transition fluxes from the wild-type state to a targeted production state. In highly connected metabolic networks, flux distribution is controlled by multiple enzymes in the network, often indirectly linked to the pathway of interest. As such, the determination of the enzyme targets for over or under expression in the metabolic network has been

challenging. The goal of this work is to develop a method to achieve this task without a complete kinetic model. Stoichiometry-based methods have been developed to handle the effects of gene knockouts that change stoichiometry. Since expression tuning does not alter metabolic network stoichiometry, such approaches cannot give an appropriate treatment for expression level changes. Furthermore, flux calculation or flux measurements give the result, but not the cause of flux transition. With anecdotal exceptions, the key enzymes whose expression changes cause the flux transition cannot be identified from the flux map. On the other hand, the traditional Metabolic Control Analysis (MCA) requires either a complete kinetic model or extensive experimentation to determine the control coefficients, which only give local but not long-range predictions. The method postulates that the most efficient way to achieve the flux change is to alter a minimal amount of enzyme levels while allowing metabolite changes to help drive the flux transition. With this assumption and a data-based sampling of reversibility, the enzymes that need to be tuned to achieve a particular flux for production can be predicted.

Finally, the derepression of the nitrogenase complex and subsequent production of hydrogen is not the only mechanism by which RubisCO deletion strains are capable of maintaining redox poise. We have isolated adaptive mutant strains that use other means to do this. One such mutant that we are currently examining is an *R. sphaeroides* RubisCO strain that up-regulates key enzymes involved in sulfate reduction, serine biosynthesis, and cysteine biosynthesis.

1. Joshi, H. M., and F. R. Tabita. A global two component signal transduction system that integrates the control of photosynthesis, carbon dioxide assimilation, and nitrogen fixation. *Proc. Natl. Acad. Sci. USA.* **93** (1996), 14515–14520
2. Tichi, M. A., and Tabita, F.R. Maintenance and control of redox poise in *Rhodobacter capsulatus* strains deficient in the Calvin-Benson-Bassham pathway. Arch. Microbiol. **174** (2000) 322-333.
3. Tichi, M. A. and Tabita, F.R. Interactive control of *Rhodobacter capsulatus* redox balancing systems during phototrophic metabolism. J. Bacteriol. **183** (2001) 6344-6354.
4. 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.

Integrated Nondestructive Spatial and Chemical Analysis of Lignocellulosic Materials during Pretreatment and Bioconversion to Ethanol

Gary Peter* (gfpeter@ufl.edu), Steve Blackband, Lonnie Ingram, Richard Yost

University of Florida, Gainesville, FL

Bioconversion of lignocellulosic biomass to ethanol involves four main operations: pretreatment, hydrolysis or saccharification, fermentation, and distillation. While improvement in each of these steps is required to make bioethanol more competitive with nonrenewable fuels, pretreatment and saccharification of the biomass are critical first steps. A number of pretreatment methods increase the yield of fermentable sugars over non-pretreated biomass. However, our fundamental knowledge of what changes pretreatment causes in biomass architecture, permeability, cellular structure, subcellular distribution, composition and organization of polymers in lignocellulosic cell walls and how these changes promote or inhibit digestion by cellulases remains limited. This limited knowledge exists despite rapid advancements in noninvasive, quantitative imaging technologies from other fields of research. Our *long-term goal* is to develop a quantitative structural model for changes that occur in the organization and chemical composition of plant biomass during pretreatment, enzymatic degradation and bioconversion to ethanol or other products. The *objectives* of this proposed work are to 1) use advanced high resolution magnetic resonance microscopy (MRM), micro and nano x-ray computed tomography (x-ray CT), electron microscopy and imaging mass spectrometry (IMS) to quantify changes in the architecture, porosity, permeability, surface area, pore size, interconnectivity, chemical organization and composition of corn stover and particularly Populus and pine wood chips during pretreatment and enzymatic degradation, and 2) integrate the results from these different quantitative imaging methods into a model for disassembly of the plant cell wall during pretreatment and bioconversion. To accomplish these objectives we will image the same biomass materials at various times during pretreatment and enzymatic degradation of cellulose. We expect to develop this model using MRM and x-ray CT images at about 5 uM resolution. The *rationale* for this proposed research is that 1) the porosity and permeability of the cell walls limit degradation and better quantification of changes in surface area, pore size, interconnectivity, porosity and permeability, spatial arrangement of cellulose, hemicellulose and lignins during pretreatment and digestion will be essential information for researchers developing improved pretreatment and enzymatic degradation methods, 2) knowledge about the structures that are recalcitrant or more slowly degraded are needed to develop faster and more efficient degradation processes, and 3) such information will be valuable to plant geneticists whose aim is to alter the structure and composition of plant biomass to make them more amenable to bioconversion processes.

Development of Computational Tools for the Analysis and Redesigning Metabolic Networks, (Award No. DE-FG02-05ER25684), Costas D. Maranas, Penn State, PI and Anthony Burgard, Genomatica, Inc. (coPI).

Patrick F. Suthers, Vinay Satish Kumar, Young Jung Kim, and Costas D. Maranas*

Department of Chemical Engineering, The Pennsylvania State University, University Park, PA 16802

The rapid progress in the generation of models of metabolism for an increasing number of organisms brings to the forefront two research questions. How can computations be leveraged to automatically assess the quality of the models and subsequently correct them by using the full complement of available experimental data ? How can we optimally combine labeling experimens with external flux measurements to elucidate the maximum number of fluxes in the model ? In this poster, we highlight progress towards both of these objectives.

Starting with the first question, our systematic metabolic model development follows four key steps. These include 1) identification of biotransformations using homology searches 2) assembly of reaction sets into a genome-scale metabolic model 3) network analysis and 4) evaluation and improvement of model performance when compared against *in vivo* data. A key challenge in the automated generation of genome-scale reconstructions is the elucidation of gaps and the subsequent generation of hypotheses to bridge them (Step 3). We have proposed optimization-based procedures to identify and eliminate network gaps in these reconstructions that we have named GapFind and GapFill. First we identify the metabolites in the metabolic network reconstruction that cannot be produced or consumed under any uptake conditions and subsequently we identify the reactions from a customized multi-organism database that restores the connectivity of these metabolites to the parent network using four mechanisms. This connectivity restoration is hypothesized to take place through four mechanisms: a) reversing the directionality of one or more reactions in the existing model, b) adding reaction from another organism to provide functionality absent in the existing model c) adding external transport mechanisms to allow for importation of metabolites in the existing model, and d) restoring flow by adding intracellular transport reactions in multi compartment models. To address the evaluation and refinement of the model performance when compared to *in vivo* data (Step 4), we have developed an optimization-based procedure (i.e., GrowMatch) to identify mismatches and propose corrective hypotheses for the model. The GrowMatch procedure identifies mismatches between the *in silico* and *in vivo* growth predictions of gene deletion mutants. GrowMatch generates hypotheses for minimally perturb the model (i.e., adding or removing functionalities) so as consistency with all experimental observations is achieved.

We have applied GapFind, GapFill amd GrowMatch to existing metabolic reconstructions such as the most recent *E. coli* and *S. Cerevisiae* models revealing many gaps and ways of restoring them. We have also deployed these tools during the construction phase of a metabolic model for *Mycoplasma genitalium.* With a genome size of ~580 kb and approximately 480 protein coding regions, *Mycoplasma genitalium* is one of the smallest known free-living organisms.

The reduced genomic content of *M. genitalium* has led researchers to suggest that the molecular assembly represented by this organism may be a close approximation to the minimal set of genes required for bacterial growth. GapFind and GapFill were used to connect three metabolites by reversing the directions of three reactions, 21 metabolites by adding 12 external reactions, and an additional 12 metabolites by adding uptake reactions. Using GrowMatch, we were able to improve the percent agreement between gene-essentiality studies *in silico* as compared to in vivo experiments from 74% to 82% We have also used the model to guide the design the composition of a defined medium.

Development of a complete and balanced metabolic model does not necessarily mean that all internal fluxes are known or changes in metabolism in response to environmental or genetic perturbations can be predicted. The gold standard in elucidating fluxes in metabolic models is the use of ^{13}C isotopic label tracing experiments. In this effort, we constructed a large-scale *E. coli* isotopomer mapping model including 393 fluxes and 214 metabolites leading to over 17,000 isotopomers. The model accounts for balances on cofactors such as ATP and NADH as well as the electron transport chain, full amino acid biosynthesis and degradation, and a detailed biomass equation. Experimental results of flux elucidation are presented for an *Escherichia coli* strain engineered to produce amorphadiene, a precursor to the anti-malarial drug artemisinin. These include a statistical analysis of fluxes determined for the system such as the minimal and maximal values of the fluxes given measurement noise.

One of the important considerations for analysis of the isotope model is how isotope measurements impact the elucidation of fluxes in large-scale metabolic reconstructions. This identifiability problem in metabolic flux analysis (MFA) with isotopic considerations is very difficult as isotopic balances yield nonlinear constraints. Here, we employ an integer programming (IP) framework for the mathematical analysis of metabolic pathways to answer this question. By using a degrees of freedom based optimization method it is possible to exhaustively identify all combinations of isotope labeling experiments and flux measurement that completely resolve all flux values in the network. This approach results in an integer linear programming formulation while accounting for the case of partial measurements (e.g., when only some fragments are measured). These measurements, consisting of both partial or full isotope state determination, were assigned relative costs that allow the experimentalist to select the measurements that will be both sufficient and economical. The proposed framework has been tested on well-studied small demonstration examples. We present benchmarks of the proposed framework by applying it to medium-scale metabolic networks of *E. coli* and by revisiting our large-scale *E. coli* model to exhaustively identify all measurements options.

1,2,4-Butanetriol trinitrate (BTTN) is manufactured by the nitration of 1,2,4-butanetriol (BT). The challenges associated with chemical synthesis of BT will be discussed along with the creation of a biosynthetic pathway that allows a single microbe to catalyze the conversion of D-xylose into D-BT. Central to this created pathway is the discovery of the ability of *Escherichia coli* to catabolize D-xylonic acid and the role that the enzyme D-xylonate dehydratase plays in this catabolism. The BT biosynthetic pathway was assembled in an *E. coli* host and begins with oxidation of D-xylose to D-xylonic acid. D-Xylonate dehydrogenase, which is heterologously expressed in an *E. coli* host from the *Caulobacter crescentus xdh* locus, is recruited for this purpose. Two xylonate dehydratases encoded by *xjhG* and *yagF* loci, which were discovered to be native to *E. coli*, catalyze the conversion of D-xylonic acid into 3-deoxy-D-*glycero*-pentulosonic acid. Decarboxylation of 3-deoxy-D-*glycero*-pentulosonic acid to form 3,4-dihydroxy-D-butanal is mediated by heterologously expressed *mdlC* isolated from *Pseudomonas putida*. Final reduction of 3,4-dihydroxy-D-butanal to BT is catalyzed by an alcohol dehydrogenase native to the BT-synthesizing *E. coli*. BTTN is more stable than nitroglycerin and mixes effectively in a solvent-free process with nitrocellulose. These characteristics make BTTN an ideal replacement for nitroglycerin and a useful plasticizer in single-stage rocket motors.

Mechanisms for Transnational Coordination and Harmonization of Nanotechnology Governance

Gary E. Marchant, Kenneth W. Abbott, and Douglas J. Sylvester

Sandra Day O'Connor College of Law, Arizona State University, Tempe, AZ

Nanotechnologies are a rapidly developing set of emerging technologies being pursued by industry and governments around the world. While these technologies promise many benefits, they will also inevitably create some risks, and regulatory agencies in numerous countries are now considering regulatory oversight approaches for nanotechnology. This project is investigating models and approaches for coordinating or harmonizing international regulation of nanotechnology. The first step in the project is to create a publicly-accessible online database of transnational, national, and sub-national regulatory activities specific to nanotechnology. Examples of entries of this database, which is scheduled to go "live" by June 2008, will be presented. The second part of the project will be to identify and analyze nine different regulatory models for transitional oversight of nanotechnologies. The nine models that will be explored will be listed and briefly described.

Discovery of Novel Machinery for Lactate Utilization by *Shewanella oneidensis* MR-1

Grigoriy Pinchuk[1], Dmitry Rodionov[2*](rodionov@burnham.org), Chen Yang[2], Xiaoqing Li[2], Samantha Reed[1], Oleg Geydebrekht[1], Margaret Romine[1], Alex Beliaev[1] and **Andrei Osterman**[2,3] (osterman@burnham.org)

*Presenting author

[1]Pacific Northwest National Laboratory, Richland, Washington 99352; [2]Burnham Institute for Medical Research, La Jolla, California 92037; and [3]Fellowship for Interpretation of Genomes, Burr Ridge, Illinois 60527

Lactate is one of the major fermentative metabolism products for many microorganisms and is the most frequently used substrate for experimental studies of respiratory metabolism in *Shewanella oneidensis* MR-1. Consequently, the metabolism of lactate is a key component of the systems-level conceptual model (under development by the Shewanella Federation) linking electron transfer networks and central/peripheral carbon metabolism pathways of MR-1. Whereas physiological data has demonstrated a robust growth of *S. oneidensis* on both D- and L- forms of lactate, its genome does not contain orthologs of classical lactate dehydrogenases (LDH) such as D-LDH (gene *dld*) or L-LDH (gene *lldD*) of *E. coli.* We report here the discovery of a novel D- and L-lactate oxidative utilization machinery identified via a comparative genomic reconstruction of *S. oneidensis* MR-1 metabolism combined with physiological, genetic, and biochemical studies.

A hypothetical FeS-containing protein encoded by SO1521 was deemed a candidate for the missing D-LDH based on its presence in the putative operon with an ortholog of lactate permease (SO1522) and its remote homology with the FAD-containing D-LDH from yeast. This prediction was verified by analysis of a SO1521 targeted gene deletion mutant and by genetic complementation of *Escherichia coli* Δ*dld* mutant with a plasmid encoding SO1521. A detailed reconstruction and comparative analysis of *lactate utilization* subsystem including associated operons and regulons, across hundreds of bacterial genomes integrated in The SEED genomic platform (http://theseed.uchicago.edu/FIG/subsys.cgi) led to a conjecture that an adjacent three-gene operon SO1518-SO1520 comprised a previously uncharacterized enzymatic complex for the utilization of L-lactate. Two genes of this operon, SO1519 and SO1520 (previously annotated as a hypothetical FeS oxidoreductase and a ferredoxin-like protein, respectively) appear to form a core of this complex conserved in many divergent bacteria (e.g., uncharacterized operons *ykgEF* in *E. coli* and *yvfVW* in *Bacillus subtilis*). This prediction was validated by assay of targeted gene deletions in *S. oneidensis* and by genetic complementation and testing of the L-LDH enzymatic activity in *E. coli* Δ*lld* mutant overexpressing the SO1518-SO1520 operon. Furthermore, the inability of only two of 19 *Shewanella* sp. with completely sequenced genomes to grow with lactate as sole carbon source are consistent with the results of our comparative genome analysis of these species. These findings, in addition to the identification of previously unknown genes involved in lactate utilization in most *Shewanella* species, broadly impact our knowledge of this important aspect of carbon and energy metabolism in many other bacteria. Additional experiments are in progress to elucidate the details of the novel L-LDH complex in *S. oneidensis* and other species. This project is a component of the *Shewanella* Federation and, as such, contributes to the overall goal of applying the genomic tools to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus.

Single-Cell Genomics of *Prochlorococcus*

Sébastien Rodrigue*[1] (s_rod@mit.edu), Rex Malmstrom[1], Matthew Henn[2], and **Sallie W. Chisholm**[1]

[1] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 15 Vassar street, Cambridge, Massachussetts 02139, USA.
[2] The Broad Institute, 7 Cambridge Center,Cambridge, MA 02142, USA.
*Presenting author

Complete genome sequencing of individual microbes is rapidly coming a reality through the combination of randomly-primed whole genome amplification and next generation sequencing platforms. One of the critical requirements in developing robust single-cell genomics pipelines is the elimination of background DNA from the sample before amplifying the genome of the single cell. This is particularly challenging when trying to study single microbial cells from the wild, such as the ubiquitous marine cyanobacterium *Prochlorococcus*. To overcome this challenge, we are developing a flow-cytometry based approach to remove single *Prochlorococcus* cells from the high concentrations of free-DNA normally found in the seawater. Our high throughput approach entails sorting individual cells into separate wells of a 96-well plate, amplifying extracted DNA with phi29 polymerase, and PCR screening of amplified genomic material using diagnostic primer sets. This pipeline not only provides products for whole genome sequencing of selected cells, but also enables multi-locus sequence analysis of natural populations without the need for isolation of different strains. To validate the method, we first sorted individual cells of the cultured *Prochlorococcus* strain MED4, and prepped them for sequencing using Solexa. We obtained genome coverage >80%, although there was substantial unevenness of the amplified genomic material. We are also initiating an effort to sort and amplify single *Prochlorococcus* cells collected in the tropical South Pacific ocean, where we expect to find interesting genomic variants.

Use of the D. radiodurans Repair System as a Possible Method for Assembling Synthetic Chromosomes

Sanjay Vashee*, Ray-Yuan Chuang*, Christian Barnes, Michael Montague, Hamilton O. Smith, and **J. Craig Venter** (PI)

*Co-Investigators

The J. Craig Venter Institue, Rockville, MD 20850

A major goal of our Institute is to rationally design synthetic microorganisms that are capable of carrying out the required functions. One of the requirements for this effort entails the packaging of the designed pathways into a cohesive genome. One of our approaches to this problem was to develop an efficient in vitro DNA repair system based upon *Deinococcus radiodurans* (Dr). This bacterium was selected because it has the remarkable ability to survive exposure to doses as high as 15, 000 Gy of ionizing radiation, which are otherwise lethal to almost all other organisms. Although hundreds of double-strand breaks are created during this exposure, Dr is able to accurately restore its genome without evidence of mutation within a few hours after exposure, strongly suggesting that the bacterium has a very efficient DNA repair mechanism.

Since the mechanism of DNA repair is not yet well understood in Dr, we undertook several general approaches to study this phenomenon. First, we sought to establish an endogenous extract that was capable of carrying out DNA repair. This extract can then be fractionated to isolate and purify all proteins that perform this repair. Second, we made use of information gathered from the sequenced genome. For example, homologues of *E. coli* DNA repair proteins, such as recA, recD and ruvA, etc. are present in Dr. Thus, we sought to characterize all these proteins both biochemically and genetically to probe their role in DNA repair. Representative results based on these studies are presented in the poster.

Engineering Isobutanol Production in *E. Coli*

Thomas Buelter* (tbuelter@gevo.com), Renny Feldman, Peter Meinhold

Gevo, Inc., Pasadena, CA

Advanced biofuels, which include isobutanol, will deliver the performance of gasoline without the environmental impact and these biofuels will reduce our dependency on foreign oil. Isobutanol has a higher energy content per gallon than many first generation biofuels, it does not absorb water and can be transported through the existing oil and gas distribution infrastructure. Isobutanol can be used in gas-powered vehicles without modification or blending.

Gevo has licensed technology for the production of isobutanol from metabolically engineered *E. coli* cells from the University of California Los Angeles where it was developed by Prof. James Liao. In order to make isobutanol competitive on the fuels market it has to be produced with high yield. We are working on recombinant microorganisms that are engineered to convert biomass into isobutanol without byproducts. We use strain engineering and pathway engineering to increase the yield of the isobutanol production strain. First bottlenecks and limitations of the host strain are identified. Then metabolic engineering strategies focus on directing the carbon flux from the carbon source to isobutanol completely. Pathways that produce byproducts are deactivated and the flux to isobutanol is maximized.

Creating a Pathway for the Biosynthesis of 1,2,4-Butanetriol

John Frost

1,2,4-Butanetriol trinitrate (BTTN) is manufactured by the nitration of 1,2,4-butanetriol (BT). The challenges associated with chemical synthesis of BT will be discussed along with the creation of a biosynthetic pathway that allows a single microbe to catalyze the conversion of D-xylose into D-BT. Central to this created pathway is the discovery of the ability of *Escherichia coli* to catabolize D-xylonic acid and the role that the enzyme D-xylonate dehydratase plays in this catabolism. The BT biosynthetic pathway was assembled in an *E. coli* host and begins with oxidation of D-xylose and D-xylonic acid. D-Xylonate dehydrogenase, which is heterologously expressed in an *E. coli* host from the *Caulobacter crescentus xdh* locus, is recruited for this purpose. Two xylonate dehydratases encoded by *xjhG and YagF* loci, which were discovered to be native to *E. coli,* catalyze the conversion of D-xylonic acid into 3-deoxy D-*glycero*-pentulosonic acid. Decarboxylation of 3-deoxy-D-*glycero*-pentulosonic acid to form 3,4-dihydroxy-D-butanal is mediated by heterologously expressed *mdlC* isolated from *Pseudomonas putida.* Final reduction of 3,4-dihydroxy-D-butanal to BT is catalyzed by an alcohol dehydrogenase native to the BT-synthesizing *E. coli.* BTTN is more stable than nitroglycerin and mixes effectively in a solvent-free process with nitrocellulose. These characteristics make BTTN an ideal replacement for nitroglycerin and a useful plasticizer in single-stage rocket motors.

Applications of High Throughput Solution X-ray Scattering (SXS): Progress Toward Proteomic Scale Structural Biology

John Tainer

A core aim of the MAGGIE (Molecular Assemblies, Genes, and Genomics Integrated Efficiently) program project is to develop technologies for proteomic scale visualization of macromolecular structure. While determining function from structure remains a challenge, structure aids in understanding how macromolecules function. Detailed macromolecular structure determination from X-ray crystallography or NMR has provided a broad and deep survey of soluble biomolecules. Several techniques, both experimental and computational, exploit this information to provide significant insight into and prediction of structures which have not been probed by these often challenging atomic resolution techniques. Solution X-ray scattering (also known as SAXS: small angle X-ray scattering)has matured to provide structures at a 10Å resolution. This resolution is sufficient to elucidate a great deal of the architecture of macromolecules, how they interact and exchange products along a pathway. SAXS may be powerfully combined with information from atomic resolution and computational structure prediction methods. Here we detail our highthroughput data collection, data analysis, and data storage programs and strategies. Full SAXS data collection (including collection of buffer blanks and a 3 fold serial dilution) on a macromolecule of interest is complete in under 10 minutes. A computer processor limited data analysis tree has been developed requiring minimal human intervention. Our web accessible data storage utility BioIsis (www.bioisis.net) allows direct access to data at all stages of analysis. We demonstrate this strategy on 16 protein from pyrococus furiosis previously prepared for a crystallographic based structural genomic effort.

AGENDA

Sunday Evening, February 10, 2008

5:00-8:00 Registration and Poster Set Up
6:00-8:00 Mixer

Monday, February 11, 2008

7:00-8:00 Continental Breakfast

8:00-9:00 **Welcome, State of Genomics:GTL Program** Sharlene Weatherwax

9:00-11:30 **Plenary Session: DOE Bioenergy Research Centers** Moderator: John Houghton

9:00-9:45 Tim Donohue - University of Wisconsin
The Great Lakes Bioenergy Research Center

9:45-10:30 Martin Keller - Oak Ridge National Laboratory
The Bioenergy Science Center

10:30-10:45 Break

10:45-11:30 Jay Keasling - Lawrence Berkeley National Laboratory
The Joint Bioenergy Institute

11:30-12:00 **Joint Genome Institute Update** Eddie Rubin

12:00-2:00 Lunch

2:00-5:00 **Concurrent Breakout Sessions**

2:00-5:00 **Breakout Session 1: Metabolic Engineering Working Group (MEWG): Interagency Conference on Metabolic Engineering** Moderator: Fred Heineken, National Science Foundation

This breakout session, conducted jointly between the DOE Genomics:GTL program and the MEWG Inter-Agency Conference on Metabolic Engineering 2008, addresses how metabolic engineering facilitates the development of the biorefinery approach to chemical production. Biorefineries are a major focus of the drive to replace fossil fuels with sustainable biologically derived ones. The same biological feedstocks that can be used for fuel production can also be the starting materials for other value added chemical products, from acids to polymers. Biorefineries have the potential to produce multiple product lines at the same plant. To do so, however, often requires metabolic engineering of the organisms involved in the various production stages. This session, therefore focuses on work that addresses some of the issues associated with such chemical production, from modification of organisms for fuel and value added chemical production to techniques for measuring and manipulating the necessary metabolic networks.

2:00-2:25 Christina Smolke, Caltech
Foundational advances in RNA engineering applied to control biosynthesis

2:25-2:50 Dhinakar Kompala, University of Colorado, Boulder
Kinetic modeling of metabolically engineered Zymomonas mobilis to maximize ethanol production from a mixture of glucose and xylose

2:50-3:15 Terry Papoutsakis, University of Delaware
Development of tolerant and other complex phenotypes for biofuel production

3:15-3:40 Break

3:40-4:05 Roger Ely, Oregon State University
Metabolic engineering for maximal, 24-hour production of hydrogen gas by wild-type and mutant strains of Synechocystis PCC 6803

4:05-4:30 John Frost, Michigan State University/ Frost Chemical Laboratory
Supplanting the chemical legacy of Alfred Nobel: Creation of a pathway for the biosynthesis of 1,2,4-butanetriol

4:30-5:00 Group Discussion

2:00-5:00 **Breakout Session 2: Nanoscience Technology**

Moderator: Sharlene Weatherwax
Organizers:
Ron Zuckermann – Molecular Foundry, LBNL
Mike Simpson – Center for Nanophase Materials Science, ORNL

This tutorial will highlight opportunities bridging between nanoscience technology and genomics research. In addition to describing the specific missions of the Molecular Foundry at LBNL and the Center for Nanophase Materials Sciences (CNMS) at ORNL, presentations will illustrate how members of the Genomics:GTL community can work with these unique DOE scientific user facilities.

2:00-3:15 Ron Zuckerman – Lawrence Berkeley National Laboratory
Carolyn Pearce – University of Manchester
Joseph Mougous – University of Washington
Kent Kirshenbaum – New York University
David Robinson – Sandia National Laboratory

3:15-3:30 Break

4:00-5:00 Mike Simpson – Oak Ridge National Laboratory
Mike McCollum – Miami University
Mitch Doktycz – Oak Ridge National Laboratory
Gary Sayler – University of Tennessee

2:00-5:00 **Breakout Session 3: Methods for Studying Proteins & Protein Complexes**

Moderator: Arthur Katz
Organizer: Jim Bruce, Washington State University

Comprehensively analyzing the proteins and molecular complexes that perform life's most essential functions presents many challenges due to their large number, biochemical variations and dynamic nature. In this session, presentations will be made on a variety of advanced analytical approaches to the study of protein-protein interactions and molecular complexes. These will be followed by an open discussion period focusing on the application of new methodologies to the characterization complex molecular interactions.

2:00-2:25 Liang Shi – Pacific Northwest National Laboratory
Investigation of Protein-protein Interactions Related to Salmonella typhimurium Pathogenesis: in vivo cross-linking with formaldehyde, tandem affinity purification and mass spectrometry identification

2:25-2:50 Gareth Butland – Lawrence Berkeley National Laboratory
Title To Be Announced

2:50-3:15 Gary Siuzdak – Scripps Research Institute
A Novel Approach Toward Characterizing Protein-Metabolite Interactions

3:15-3:40 Greg Hurst – Oak Ridge National Laboratory
Analysis of Protein Complexes at the Genomics:GTL Center for Molecular and Cellular Systems

3:30-4:00 Break

4:00-5:00 Group Discussion

5:00-8:00 Poster Session Salon ABCD

Tuesday, February 12, 2008

7:00-8:00 Continental Breakfast

8:00-9:00 **Keynote Speaker - Jo Handelsman, University of Wisconsin - Madison**
Molecules of War & Conversation and the Soil Metagenome

9:00-9:30 Data Management & Sharing Policy – Susan Gregurick & Jim Fredrickson

9:30-10:00 Break

10:00-12:00 **Plenary Session: Advances in Genomic Technology** Moderator: Marvin Stodolsky

10:00-10:40 Daniel Gibson - J. Craig Venter Institute
Mycoplasma Genome Synthesis and Transplantation: Progress on Constructing a Synthetic Cell

10:40-11:20 George Church - Harvard University
Genome-wide Reprogramming and Accelerated Evolution

11:20-12:00 Daphne Preuss - University of Chicago, Chromatin Inc.
Developing Synthetic Chromosomes for Crops: Applications for Agriculture and Energy

12:00-2:00 Lunch

2:00-5:00 **Concurrent Breakout Sessions**

2:00-5:00 **Breakout Session 4: Advanced Characterization and Imaging of Lignocellulose Materials** Moderator: Roland Hirsch

Lignocellulosic materials are difficult to break down into components that are easy to process into fuels and other products. They also are very difficult to characterize sufficiently to understand in real time how they are affected by physical, chemical and biological treatments. Yet these materials are the starting point for much planned research in the Genomics:GTL program. This session will provide an overview of four analytical technologies and how they may overcome the obstacles to characterizing lignocellulosic materials. The presentations will be by scientists who have recently initiated research projects to study the application of the technologies to GTL-relevant materials. The focus will be on the capabilities of the technologies and the session will be appropriate for attendees unfamiliar with them.

2:00-2:25 Barbara Evans - Oak Ridge National Laboratory
Dynamic Visualization of Lignocellulose Degradation by Integration of Neutron Scattering Imaging and Computer Simulation

2:25-2:50 Gary Peter - University of Florida
Title To Be Announced

2:50-3:15 Hoi-Ying Holman - Lawrence Berkeley National Laboratory
Synchrotron Infrared (SIR) SpectroMicroscopy of Living Microbial Cells

3:15-3:40 Paul Bohn - Notre Dame University
Three-Dimensional Spatial Profiling of Lignocellulosic Materials by Coupling Light Scattering and Mass Spectrometry

3:40-4:00 Break

4:00-5:00 Group Discussion

2:00-5:00 **Breakout Session 5: Microbial Growth Technologies for Systems Biology** Moderator: Joe Graber

Understanding the systems biology of microorganisms requires not only advanced molecular tools but also increasingly sophisticated cultivation techniques that allow subtle manipulation of environmental variables and microbial growth states. This session will focus on advanced cultivation approaches used to study and model regulatory control of complex behaviors by microbes and microbial communities. A series of presentations will be followed by an open discussion period to discuss new methodologies, identify technical hurdles, and exchange ideas on cultivation strategies relevant to systems biology research.

2:00-2:25 Jeff McLean, J. Craig Venter Institute
Cultivation and Analytical Approaches for Systems Biology of Biofilms

2:25-2:50 Mary Lidstrom, University of Washington
Coupling Function to Genomics in Microbial Communities Via Single-Cell Analysis

2:50-3:15 Nitin Baliga, Institute for Systems Biology
Culturing Technologies for Constructing Predictive Systems- Scale Models of Microbial Behavior

3:15-3:40 Dave Emerson, Bigelow Research Laboratory
The Way Life Should Be: Using Gradients to Capture Microbes and Study Their Behavior

3:30-4:00 Break

4:00-5:00 Group Discussion

2:00-5:00 **Breakout Session 6: Web-Based Resources for Microbial Genomics** Moderators: Dan Drell and Sharlene Weatherwax

Recent advances in genome and metagenome sequencing technology have yielded not only exciting new insights on the fundamental properties of organisms and communities, but also an intimidating volume of data that is increasingly difficult to organize and study. In this session, tutorials will be provided on three web-based resources that facilitate data management and provide bioinformatics tools for the analysis of large genomic and metagenomic data sets. In addition to providing overviews of available resources, these presentations will focus on the practical application of these resources to demonstration sets of common data.

2:00-2:30 Microbes Online

Paramvir Dehal, Lawrence Berkeley National Laboratory

2:30-3:30 Integrated Microbial Genomes (IMG)
Phil Hugenholtz, Joint Genome Institute

3:30-4:00 Break

4:00-5:00 CAMERA Metagenomics Database
Rekha Seshdari, J. Craig Venter Institute

5:00-8:00 Poster Session Salon ABCD

Wednesday, February 13, 2008

7:00-8:00 Continental Breakfast

8:00-10:00 **Plenary Session: Protein Interactions, Complexes, & Networks**
Moderator: Arthur Katz

Michelle Buchanan - Oak Ridge National Laboratory
8:00-8:40 *The Center for Molecular and Cellular Systems: Biological Insights from Large Scale Protein-Protein Interaction Studies*

Mark Biggin - Lawrence Berkeley National Laboratory
8:40-9:20 *High Throughput Identification and Characterization of Protein Complexes in Desulfovibrio vulgaris*

John Tainer - Lawrence Berkeley National Laboratory
9:20-10:00 *Molecular Assemblies, Genes, and Genomics Integrated Efficiently (MAGGIE): Integrating Combined Methods and Comparative Systems to Connect Genes to Functional Networks*

10:00-10:30 Break

10:30-11:00 Zan Luthey-Schulten, University of Illinois at Urbana-Champaign
Title To Be Announced

11:00-11:30 Rick Stevens, Argonne National Laboratory
Presentation on the Joint Report BERAC/ASCAC Joint Report on Computational Modeling in the Genomics:GTL Program

11:30-12:00 Closing Remarks