

# Systems Biology Strategies and Technologies for Understanding Microbes, Plants, and Communities

## Analytical Strategies for the Study of Plants, Microbes, and Microbial Communities

# 176

### Towards an Understanding of Proteins that Govern the Structure and Function of *Synechocystis* sp. PCC6803 and Multiple *Cyanotheca* Strains

Stephen J. Callister,<sup>2\*</sup> Jon M. Jacobs,<sup>2</sup> Jana Stöckel,<sup>4</sup> Jason E. McDermott,<sup>3</sup> Lee Ann McCue,<sup>3</sup> Carrie D. Nicora,<sup>2</sup> Ljiljana Paša-Tolić,<sup>1</sup> Louis A. Sherman,<sup>6</sup> Himadri B. Pakrasi,<sup>4,5</sup> and David W. Koppenaal<sup>1,2</sup> (david.koppenaal@pnl.gov)

<sup>1</sup>Environmental Molecular Sciences Laboratory and <sup>2</sup>Biological and <sup>3</sup>Computational Sciences Division, Pacific Northwest National Laboratory, Richland, Wash.; <sup>4</sup>Depts. of Biology and Energy, <sup>5</sup>Environmental and Chemical Engineering, Washington University, St. Louis, Mo.; and <sup>6</sup>Dept. of Biological Sciences, Purdue University, West Lafayette, Ind.

**Project Goals:** The primary goal of this project is to apply a systems biology approach to understand the network of genes and proteins that govern the structure and function of *Cyanobacteria*. These microorganisms make significant contributions to harvesting solar energy, planetary carbon sequestration, metal acquisition, and hydrogen production in marine and freshwater ecosystems. *Cyanobacteria* are also model microorganisms for studying the fixation of carbon dioxide through photosynthesis at the biomolecular level. Importantly, this project addresses critical U.S. Department of Energy science needs, provides model microorganisms to apply high-throughput biology and computational modeling, and takes advantage of EMSL's experimental and computational capabilities.

In the initial phase of this project, proteomics characterizations of *Synechocystis* sp. PCC6803 and *Cyanotheca* sp. ATCC 51142 were performed to improve our understanding of the proteome makeup of these model organisms under normal and perturbed growth conditions. We have developed the most complete quantitative proteome analysis of *Synechocystis* sp. PCC6803 under various critical environmental perturbations applying a high sensitivity mass spectrometry approach spanning 33 physiological conditions. The resulting proteome dataset consists of 22,318 unique peptides,

corresponding to 2,369 unique proteins, covering 65% of the predicted proteins. Quantitative analysis of changes in protein abundance under environmental perturbations has led to the identification of the key proteins required for the maintenance of cellular fitness necessary for cell survival.

We also examined the impact of diurnal rhythms on the protein level of *Cyanotheca* 51142. We identified a total of 3,616 proteins with high confidence, which accounts for ~68% of the predicted proteins based on the completely sequenced *Cyanotheca* 51142 genome. About 77% of identified proteins could be assigned to functional categories. Quantitative proteome analysis uncovered that ~3% of the proteins exhibit oscillations in their abundance under alternating light-dark conditions. The majority of these cyclic proteins are associated to central intermediary metabolism, photosynthesis as well as biosynthesis of cofactors. Our data also suggest that diurnal changes in activities of several enzymes are mainly controlled by turnover of related cofactors and key players but not entire protein complexes.

While *Cyanotheca* sp. ATCC 51142 continues to represent a model organism for proteomics investigations, six additional *Cyanotheca* either have finished, or draft genome sequences. This number of strains, having genome sequences, allows for comparison of *Cyanotheca* at the level of the core genome and core proteome. While the core genome predicts the common phenotype of *Cyanotheca*, the core proteome represents the actual protein phenotype characteristic of *Cyanotheca*. As such, all strains were cultured in the laboratory under growth conditions best representing their natural environments and proteins extracted from cells have been analyzed by LC-MS/MS. Results from over 460 LC-MS/MS analyses have been used to develop proteomics databases for strains PCC8801, PCC8802, PCC7424, PCC7425, PCC7822, and ATCC51142. An additional proteomics database for ATCC51472 is pending on the completion of a draft genome sequence for this strain. The Proteomics database for the more distantly related *Synechosystis* sp. PCC6803 was also included in this analysis to better constrain the core proteome to represent *Cyanotheca*. While the core proteome of *Cyanotheca* is composed of a large percentage of proteins involved in energy production, translation, and amino acid production, a significant portion of the core proteome is also made up of proteins having no predicted function, or only a general assigned function. Of interest was the observation of hypothetical and conserved hypothetical proteins suggesting the importance of these proteins in defining the general lifestyle of *Cyanotheca*, yet also suggesting the need for additional functional characterization of these proteins to better understand *Cyanotheca* from a systems biology perspective.

The research was performed as part of an EMSL Scientific Grand Challenge project at the W.R. Wiley Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored

by the U.S. Department of Energy's Office of Biological and Environmental Research (BER) program located at Pacific Northwest National Laboratory. PNNL is operated for the Department of Energy by Battelle.

# 177

## Deciphering Microbial Community Dynamics via Observational and Experimental 'Metatranscriptomics': Developments and Applications

**Edward F. DeLong**<sup>1</sup> (delong@mit.edu), Elizabeth Ottesen<sup>1\*</sup> (ottesen@MIT.EDU), Adrian Sharma<sup>1\*</sup> (sharma.adrian@gmail.com), Yanmei Shi,<sup>1</sup> Gene Tyson,<sup>1</sup> Frank Stewart,<sup>1</sup> Rex Malmstrom,<sup>1</sup> Sallie W. Chisholm,<sup>1</sup> Jamie Becker,<sup>2</sup> and Dan Repeta<sup>2</sup>

<sup>1</sup>Massachusetts Institute of Technology, Cambridge and <sup>2</sup>Woods Hole Oceanographic Institution, Woods Hole, Mass.

[http://openwetware.org/wiki/DeLong\\_Lab](http://openwetware.org/wiki/DeLong_Lab)

**Project Goals:** One of our central project goals is to develop and refine methods for studying microbial community gene expression in the environment, referred to here as 'metatranscriptomics'. There are a variety of diverse applications of these new metatranscriptomic methods. The approach can be used to verify that hypothetical ORFs identified in metagenomic projects are indeed transcribed and expressed. In surveys, 'observational metatranscriptomics' can be used to survey both the nature and abundance of different RNA species (including ribosomal RNA, messenger RNA, and small non-coding RNAs) existing within any given microbial community. Finally, 'experimental metatranscriptomics' can be leveraged to reveal transcriptional responses of microbial communities to environmental perturbation, as well as to discover the specific metabolic pathways involved in matter and energy cycling. In the course of developing methods and protocols for metatranscriptomics, we have explored many of these specific applications. We report here on a few examples, including the refinement of the metatranscriptomic protocols and analyses of biological and technical replicates, the discovery of new microbial small RNAs, and the identification of metabolic pathways involved in the turnover of high molecular weight dissolved organic matter (DOM) in the marine environment.

### Development and quantitative evaluation of an rRNA subtraction protocol

Metatranscriptomes generated by pyrosequencing have great potential use for describing functional processes and attributes of complex microbial communities. Meeting this potential requires protocols that maximize mRNA recovery by reducing the relative abundance of ribosomal RNA, as well as systematic comparisons to identify methodological artifacts and test for reproducibility across datasets. We

developed a protocol for subtractive hybridization of small and large subunit RNAs using sample-specific probes, that is applicable across diverse environmental samples. To test the method, rRNA-subtracted and unsubtracted transcriptomes were pyrosequenced from several different bacterioplankton communities ocean, representing ~350 Mbp of metatranscriptomic data. The new subtractive hybridization method reduced bacterial rRNA transcript abundance by 40 to 58%, increasing recovery of non-rRNA sequences up to fourfold. To evaluate this method, we established criteria for detecting sequences replicated artificially via pyrosequencing errors and identified such replicates as a significant component (6 to 39%) of total pyrosequencing reads. Following replicate removal, statistical comparisons of reference genes (identified via BLASTX to NCBI-nr) between technical replicates and between rRNA-subtracted and unsubtracted samples showed low levels of differential transcript abundance (< 0.2% of reference genes). However, gene overlap between datasets was remarkably low, with no two datasets (including duplicate runs from the same pyrosequencing library template) sharing greater than 17% of unique reference genes. These results suggest that current levels of pyrosequencing capture a small subset of total mRNA diversity, underscoring the importance of rRNA subtraction to enhance sequencing coverage across the functional transcript pool.

### Novel small RNAs revealed by metatranscriptomics

Previous metatranscriptomic studies have suggested that many cDNA sequences share no significant homology with known peptide sequences, and therefore might represent transcripts from uncharacterized proteins. We found that a large fraction of cDNA sequences detected in a metatranscriptomic datasets are comprised of well-known small RNAs (sRNAs), as well new groups of previously unrecognized putative sRNAs (psRNAs). These psRNAs mapped specifically to intergenic regions of microbial genomes recovered from similar habitats, displayed characteristic conserved secondary structures, and were frequently flanked by genes that suggested potential regulatory functions. Depth-dependent variation of psRNAs generally reflected known depth distributions of broad taxonomic groups, but fine-scale differences in the psRNAs within closely related populations suggested potential roles in niche adaptation. Genome-specific mapping of a subset of psRNAs derived from predominant planktonic species like *Pelagibacter* revealed recently discovered as well as potentially new regulatory elements. Our analyses show that metatranscriptomic datasets can reveal new information about the diversity, taxonomic distribution and abundance of sRNAs in naturally occurring microbial communities, and suggest their involvement in environmentally relevant processes including carbon metabolism and nutrient acquisition.

### Metabolic pathways of DOM turnover revealed by metatranscriptomics

To better measure and model the microbial processes associated with the turnover of DOM in the sea, we performed metatranscriptomic analyses in experimental microcosms that were amended with DOM. High molecular weight DOM from surface waters of the North Pacific Subtropical Gyre near station ALOHA was concentrated by ultrafiltration

tion using a 1 nm pore membrane filter, and added to unfiltered seawater microcosms. The twenty liter microcosms were maintained at *in situ* temperatures and light intensities, and sampled periodically over the course of a 27 hour incubation period. In conjunction with metagenomic datasets obtained at the beginning and end of the experiment, samples for metatranscriptomic analyses were collected over the time course of the experiment in both the unamended control, DOM enriched sample. Subsequent analyses revealed the timing, potential biochemical pathways, microbial species, and potential organic carbon compound intermediates associated with HMW DOM degradation. The results suggested a successional cascade of microbial species related to stepwise metabolic transformations involved in microbially mediated oxidation of DOM in the sea.

# 178

## Identification Characterization of Methanosarcinaceae Cell Surface Proteins

**R.P. Gunsalus**<sup>1,2\*</sup> (robgs@microbio.ucla.edu),  
D. Francoleon,<sup>3</sup> R. Loo,<sup>3</sup> J. Loo,<sup>1,3</sup> L. Rohlin,<sup>2</sup> U. Kim,<sup>2</sup> and  
**M. Arbing**<sup>1</sup>

<sup>1</sup>UCLA-DOE Institute of Genomics and Proteomics,  
<sup>2</sup>Dept. of Microbiology, Immunology, and Molecular  
Genetics, and <sup>3</sup>Dept. of Chemistry and Biochemistry,  
University of California, Los Angeles

<http://www.doe-mbi.ucla.edu/people>

**Project Goals:** The goal of this research project is to gain a better understanding of the microorganisms that capture, store and mobilize energy, processes that occur naturally in the Earth's biosphere. We are investigating the molecular biology and biochemistry of several model methane producing archaea and associated hydrogen producing syntrophic bacteria. As participants of anaerobic food chains, they aid in the conversion of complex plant and animal polymers to a variety of small molecular weight carbon intermediates (e.g., alcohols, short chain fatty acids, and various aromatic compounds) to hydrogen, methane and carbon dioxide. However, still lacking is a clear definition of key metabolic pathways, energy conserving complexes, and cell architectures needed to use the above compounds for methane/hydrogen end product formation. We are developing tools to study these anaerobic microorganisms to assess and model hydrogen and methane production. This includes development of community resources for mRNA enrichment/sequencing methods for transcript analysis of model microbes, application of proteomic methods to further define the unique biology, and exploratory metabolomic studies to document the metabolic intermediates in these model microbes. Development and application of such tools will allow better assessment, and modeling of these poorly understood and underutilized hydrogen and methane producing microorganisms for future exploitation.

The cell envelopes of many archaeal species contain a proteinaceous lattice termed the surface layer or S-layer. It is typically composed of only one or two abundant, often post-translationally modified proteins that self-assemble to form a highly organized cell surface-exposed array. Little is known about these proteins in any methanogenic archaean. Surprisingly, over a hundred proteins were annotated to be S-layer or surface associated components in the *Methanosarcina acetivorans* C2A and *Methanosarcina mazei* Gö1 genomes, reflecting limitations of current bioinformatics predictions. To experimentally address what proteins are present, we devised an *in vivo* biotinylation technique to affinity tag all surface-exposed proteins. This overcame several challenges in working with these fragile microorganisms. The two *Methanosarcina* species were adapted to growth under N<sub>2</sub> fixing conditions to minimize the level of free amines that would interfere with the NHS-label acylation chemistry used. A 3-phase separation procedure was then employed to isolate the intact labeled cells from any lysed-cell derived proteins. The Streptavidin affinity enrichment was followed by stringent wash to remove non-specifically bound proteins, and LC-MS-MS methods were employed to identify the labeled surface proteins. In *M. acetivorans* C2A and *M. mazei* Gö1 the major surface layer proteins were identified to be the MA0829 and MM1976 gene products, respectively. Each of the proteins were shown to exist in multiple forms by using SDS-PAGE coupled with glycoprotein-specific staining, and by interaction with the lectin, Concanavalin A. Of the less abundant surface-exposed proteins identified, the presence of all three subunits of the thermosome suggests that the archaeal chaperonin complex is both surface- and cytoplasmically-localized. The above-described techniques provide an alternative strategy to isolate and characterize cell envelope proteins in these archaea.

In related studies we are characterizing the molecular and structural properties of the above surface layer proteins. The *M. acetivorans* MA0829 protein possesses two domains of unknown function that are 78% identical and 86% similar. X-ray crystallography is being used to gain insight into this structure whereby crystallization screening has yielded crystals that diffract to 2.4 Å. Structure solution using selenomethioine-labeled protein is in progress. Finally, bioinformatics searches have revealed the distribution of related surface layer proteins in the Methanosarcinaceae and in other archaeal species.

## 179

## Coupling Function to Phylogeny via Single-Cell Phenotyping

Marina G. Kalyuzhnaya<sup>2\*</sup> (mkalyuzh@u.washington.edu), Sarah McQuaide,<sup>1</sup> Ekaterina Latypova,<sup>1</sup> Samuel Levine,<sup>1</sup> David Ojala,<sup>1</sup> Michael Konopka,<sup>1</sup> and **Mary E. Lidstrom**<sup>1,2</sup> (lidstrom@u.washington.edu)

<sup>1</sup>Depts. of Chemical Engineering and <sup>2</sup>Microbiology, University of Washington, Seattle

**Project Goals:** 1) Develop new technology for presorting functional populations and analyze them at the single cell level for both phenotypic and genomic parameters. 2) Apply this approach to populations from Lake Washington sediments to couple functional and genomic datasets at the single cell level.

Rapid advances in modern molecular methods such as the whole genome community sequencing (WGCS) approach open new ways to study microbial ecology. While application of the high-throughput sequencing could result in a blueprint of genomic content of the ecosystem of interest, generally it provides little information about ecological significance of the newly detected functions. To truly understand the role of microbes in the environment, the genomic sequences should be reconsidered in the context of physiological data. Integration of single-cell physiological measurements with genomic data in order to elucidate the functional role of yet uncultivable microbes is the major focus of our current research. Overview of our approaches is presented in Figure 1. We use respiration as a core metabolic function to describe methylotrophic capabilities of microbial cells from two natural environments: freshwater lake sediment (Lake Washington) and salt water (Saanich Inlet). Subsequently, cells tested positive for specific functions are targeted for further genomic explorations via whole genome amplification, PCR-surveys for functional genes and whole genome sequencing.

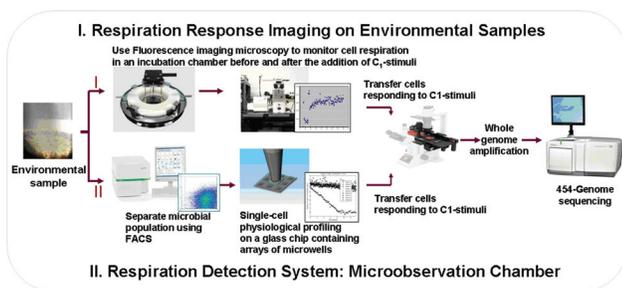


Figure 1. Proposed approaches to couple physiological function with genomics

This single cell platform is a new way to uncover or refine the function for yet uncultured members of natural microbial communities.

## 180

## Proteomics Driven Analysis of Microbes, Plants and Microbial Communities

**Mary S. Lipton**<sup>1\*</sup> (mary.lipton@pnl.gov), Stephen J. Callister,<sup>1</sup> Kristin E. Burnum,<sup>1</sup> Roslyn N. Brown,<sup>1</sup> Haizhen Zhang,<sup>1</sup> Thomas Taverner,<sup>1</sup> Margaret F. Romine,<sup>1</sup> Kim K. Hixson,<sup>2</sup> Janani Shutthanandan,<sup>1</sup> Samuel O. Purvine,<sup>2</sup> Angela D. Norbeck,<sup>1</sup> Matthew E. Monroe,<sup>1</sup> Carrie D. Nicora,<sup>1</sup> Gordon A. Anderson,<sup>1</sup> Steve Giovannoni,<sup>3</sup> Jerry Tuscan,<sup>4</sup> Phil Hugenholtz,<sup>5</sup> Stuart Levy,<sup>6</sup> Derek Lovley,<sup>7</sup> and **Richard D. Smith**<sup>1</sup>

<sup>1</sup>Biological Sciences Division and <sup>2</sup>Environmental Molecular Sciences Laboratory; Pacific Northwest National Laboratory, Richland, Wash.; <sup>3</sup>Oregon State University, Corvallis; <sup>4</sup>Oak Ridge National Laboratory, Oak Ridge, Tenn.; <sup>5</sup>DOE Joint Genome Institute, Walnut Creek, Calif.; <sup>6</sup>Tufts University, Boston, Mass.; and <sup>7</sup>University of Massachusetts, Amherst

**Project Goals:** This project employs comprehensive global and directed proteomic analyses of microbes, plants and microbial communities to enhance the scientific understanding through improved annotation of genomic sequences, elucidation of phenotypic relationships between environmentally important microorganisms, characterization of higher organisms, characterization of the metabolic activities within microbial communities, and identification of post-translationally modified proteins.

Inherent to exploiting microbial function or utilizing plants as biofuels is the detailed understanding of the physiology of the cell. These cellular functions are dictated by the proteins expressed in the cell, their localization and their modification state. This project exploits the technological and informatics advances in the proteomics pipeline at PNNL (as described in the poster by Anderson et al) to address organism-specific scientific objectives developed in conjunction with biological experts for a number of different microbes and plants. In our poster, we highlight the ability to use proteomics data for genome annotation of microbes and fungi, characterization of microbial communities, advances in the characterization of protein phosphorylation state, and the identification of new proteins important to photosynthesis, and the determination of protein localization in stem, root and leaf tissues of poplar.

Genome sequences are annotated by computational prediction of coding sequences, followed by similarity searches such as BLAST, which provide a layer of (possible) functional information. While the existence of processes such as alternative splicing complicates matters for eukaryote genomes, the view of bacterial genomes as a linear series of closely spaced genes leads to the assumption that computational annotations which predict such arrangements completely describe the coding capacity of bacterial genomes. However, proteomic experiments have shown the expression in *Pseudomonas fluorescens* Pf0-1 of sixteen non-annotated

protein-coding regions, of which **nine were antisense to predicted genes**, six were intergenic, and one read in the same direction as an annotated gene but in a different frame. The expression of all but one of the newly discovered genes was verified by RT-PCR. Few clues as to the function of the new genes were gleaned from informatic analyses, but potential orthologs in other *Pseudomonas* genomes were identified for eight of the new genes. The 16 newly identified genes improve the quality of the Pf0-1 genome annotation, and the detection of antisense protein-coding genes indicates the under-appreciated complexity of bacterial genome organization.

Unique to proteomic studies, the elucidation of post-translational modifications and protein localization lead to a richer understanding of the biological system. We have developed advanced technologies to fractionate proteins from microbial subcellular fractions and have applied this technology to mixtures of dissimilar microbes. The applications of this technology to microbial communities will result in a reduction of the sample complexity and increase characterization of the community. Heme moieties play an important role in microbial respiration, yet to date have remain recalcitrant to proteomic characterization. Application of refined separation strategies have resulted in samples enriched in heme containing proteins and thus aid in the identification of these proteins.

Proteomics characterization is also used to understand more complex systems such as plants and microbial communities. *Populus* is the fastest growing tree species in North America and has been identified as a potentially important crop species for converting plant biomass to liquid fuels. *Populus* species are broadly adapted to nearly all regions of the U.S., and hybrid clones have demonstrated 10 dry tons per acre productivity on a commercial scale. Still, improvements in growth rate, cell wall composition, drought tolerance, and pest resistance are required before this species reaches its potential as an energy crop. We have used proteomics technologies to map the protein expression patterns between root, leaf and stem tissues.

Termites degrade and thrive on lignocellulose with help from the bacterial microbiome harbored within their guts. Recent metagenomic analyses have begun to shed light on the genetic potential of the termite hindgut community, but little is known about which genes are expressed to support the symbiotic relationship. Here, we analyzed the metaproteome of the bacterial community resident in the hindgut paunch of the wood-feeding 'higher' *Nasutitermes* species and identified 886 proteins, 197 of which have known enzymatic function. Using these enzymes, we reconstructed known metabolic pathways to gain a better understanding of carbohydrate transport and metabolism, nitrogen fixation and assimilation, energy production, and amino acid synthesis in this endosymbiotic microbiome.

Additional information and supplementary material can be found at the PNNL proteomics website at <http://oberproteomics.pnl.gov/>

This research is supported by the Office of Biological and

Environmental Research of the U.S. Department of Energy. Portions of this research were performed in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the DOE's Office of Biological and Environmental Research. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute through Contract No. DE-AC05-76RLO 1830.

## 181

### Advances for High Throughput, Comprehensive and Quantitative Proteomics and Metabolomics Measurements; Enabling Systems Biology

**Gordon A. Anderson\***, Ronald J. Moore, David J. Anderson, Kenneth J. Auberry, Mikhail E. Belov, Kevin Crowell, Stephen J. Callister, Therese R.W. Clauss, Kim K. Hixson, Gary R. Kiebel, Brian L. LaMarche, Mary S. Lipton, Da Meng, Thomas O. Metz, Matthew E. Monroe, Heather M. Mottaz, Carrie D. Nicora, Angela D. Norbeck, Daniel Lopez-Ferrer, Daniel J. Orton, Ljiljana Paša-Tolić, David C. Prior, Samuel O. Purvine, Anuj Shah, Yufeng Shen, Anil K. Shukla, Mudita Singhal, Gordon W. Slys, Aleksey V. Tolmachev, Nikola Tolić, Karl Weitz, Aaron Wright, Rui Zhang, Rui Zhao, and Richard D. Smith (rds@pnl.gov)

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

**Project Goals:** This project is developing and applying new and greatly expanded quantitative, high throughput proteomics and metabolomics capabilities for studying diverse microbial systems and communities, plants, and ecosystems of increasing levels of complexity. Capability advances this past year centered on extending modified protein coverage through the integration of bottom-up and top-down measurements, the development of multiplexed activity-based proteomics, developing broad metabolome measurement coverage, and the integration of proteomics and metabolomics measurements with genomic information. In concert with other measurements and information, these developments are addressing deficiencies in the coverage of biochemical components provided by current measurement capabilities, and thus providing the detailed data and the quality needed to enable truly effective systems biology approaches. FWP number: 40601

Understanding microbial and bioenergy-related systems requires knowledge of the array of proteins and their complement of post-translational modifications, as well as knowledge of the large (and often unknown) range of metabolites and other cellular components. Among the basic challenges associated with gaining this understanding are identifying and quantifying large sets of proteins, modified proteins, and metabolites whose relative abundances

typically span many orders of magnitude, and doing so in a sufficiently high throughput manner. Compounding these analytical challenges is the largely unknown extent and nature of many protein modifications, and the high chemical and structural diversity of metabolites.

PNNL is developing and applying high throughput mass spectrometry-based measurement technologies and associated informatics tools applicable to a broad range of biological studies, many of which are presently conducted in collaboration with a number of BER Genomic Sciences researchers (see poster by M.S. Lipton et al.). Our high throughput proteomics/metabolomics analysis pipeline is based on high resolution nano-liquid chromatography separations combined with high mass measurement accuracy mass spectrometry measurements. This poster highlights several developments that build upon this foundation:

**1. Providing much broader protein coverage, including modification states to which current measurements are effectively 'blind'.** We developed new approaches that combine top-down and bottom-up measurements to extend quantitative proteome coverage to a large range of protein modification states, and that integrate measurements from targeted post-translationally modified sub-proteomes. As part of these efforts, we combined a "RePlay" chromatography method for on-line reanalysis of the separated proteome components with an ultra-fast post-column pressure digestion system to attain nearly continuous mass spectrometer utilization and simultaneous acquisition of both top-down and bottom-up proteomics data from a single analysis, which avoids many of the present ambiguities associated with data interpretation. In conjunction with this approach, we are exploiting the increased throughput of bottom-up measurements (see below, and poster by R.D. Smith et al.) to provide detailed measurements for targeted sub-proteomes. We also are commencing development of new informatics approaches to integrate these complementary data sets.

**2. Multiplexed activity-based proteomics.** To augment the more detailed proteomics measurements noted above, we are implementing measurements that directly measure enzyme *activities* rather than abundances, and thus measurements that account for changes in protein modification, structure, localization etc. The approach involves the synthesis of *in vitro* or *in vivo* multiplexed (for different activities) and isotopically coded chemical probes that can be applied simultaneously to discover and quantitatively follow enzymatic activities. The approach allows isolation, enrichment, and analysis of large sets of labeled 'signature' peptides that in turn enable protein identification, as well as provide direct and quantitative data on a large range of biological activities in any targeted biological system.

**3. Increasing '-omics' coverage by broad nanoLC measurements of the metabolome.** We have adapted the new high throughput nanoLC-ion mobility-MS platforms noted below to obtain broad and quantitative measurements of the broad range of metabolites and small molecule components of biological systems. These measurements provide much more comprehensive data sets, which are needed to support

the development of computational models for biological systems and effective systems biology approaches.

A new fast separation liquid chromatography-ion mobility-mass spectrometry platform has been developed (and several versions now implemented) that benefits all of the above efforts by providing high levels of data quality in conjunction with an order of magnitude increase in measurement throughput (see poster by R.D. Smith et al.). The information garnered from improved global coverage of protein modifications and metabolites, and obtainable with increased throughput, is expected to have a profound impact on our ability to develop computational models of biological systems.

Gaining the full benefits of these extended measurement capabilities requires a significantly different and extended computational infrastructure. Thus, we have expanded the PNNL informatics pipeline to incorporate a suite of data analysis tools, data consolidation applications, and statistical packages, as well as visualization software for data interpretation. Through the development of these new tools and the enhancement of existing tools, we have implemented a framework that will support integration of the enhanced proteomics and metabolomics data sets. This framework further supports integration of genomics data from public repositories and provide the needed infrastructure to interoperate with the GTL Knowledgebase.

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.

## 182

### A New Platform for Much Higher Throughput, Comprehensive, and Quantitative Proteomics and Metabolomics Measurements and Data Analysis

**Richard D. Smith\*** (rds@pnl.gov), **Gordon A. Anderson**, Erin S. Baker, Mikhail E. Belov, Eric Y. Choi, Kevin L. Crowell, William F. Danielson III, Yehia M. Ibrahim, Ryan Kelly, Brian L. LaMarche, Andrei V. Liyu, Divakara Meka, Da Meng, Matthew E. Monroe, Daniel J. Orton, Jason Page, David C. Prior, Thomas A. Seim, Anuj Shah, Gordon W. Slysz, Keqi Tang, and Aleksey V. Tolmachev

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

**Project Goals:** This project aims to develop and apply greatly improved capabilities for proteomics and metabolomics. In this context, we have developed and demonstrated a new fast separation liquid chromatography/ion

**mobility/mass spectrometry platform for quantitative high throughput proteomics and metabolomics measurements that achieves high levels of data quality in conjunction with an order of magnitude increase in measurement throughput. In addition to significantly improved sensitivity and extremely large data generation rates, the new platform provides the basis for effectively generating and combining data from multiple measurements to attain broader coverage of both protein modification states and chemically diverse metabolomes. The platform considerably speeds large-scale applications and thus enables previously impractical studies, e.g., of diverse microbial systems, communities, and ecosystems. FWP number: 40601**

A major challenge underlying the successful development and application of systems biology approaches is the large numbers of measurements needed to accommodate experimental constraints, e.g., derived from available sample sizes, variability in measurements, and/or practical measurement throughput limitations. Although much greater than feasible with classical approaches (using 2D-PAGE), the proteomics measurement throughput now provided by LC-MS- and LC-MS/MS-based approaches is grossly inadequate for characterizing very large numbers of samples, e.g., involving many perturbations, or spatially and/or temporally distinct samples. Metabolomics is similarly constrained, and faces additional challenges because of the broad chemical diversity of metabolites and the greater difficulties associated with identification. Additionally, sample recovery and enrichment methodologies can limit proteome and metabolome coverages.

To address these shortcomings, we developed a new platform at PNNL that demonstrates greatly improved measurement throughput, sensitivity, robustness, and quantitative capability for proteomics and metabolomics measurements in a range of biological research applications.

The new measurement platform incorporates fast multiplexed nanocapillary LC separations coupled via a greatly improved electrospray ionization interface to an ion mobility spectrometer (IMS) stage interfaced to a high speed, accurate mass, and broad dynamic range time-of-flight mass spectrometer (TOF/MS). The automated fast nanocapillary LC system incorporates high pressure LC pumps, an autosampler, and a multiplexed 4-column fluidics system. Each 10-cm-long capillary LC column is operated at 10,000 psi to provide both fast and high resolution separations. Electrospray ionization (ESI) generated ions are accumulated at the end of the second stage of a dual electrodynamic ion funnel trap before being injected into an IMS separation drift tube stage where peptide or metabolite ion separations occur on a time scale of <50 msec. To increase IMS-TOF/MS sensitivity, we developed a novel multiplexing approach that increases the number of ion injection pulses into the IMS separation stage by >30-fold, and thus the S/N levels achievable in a given analysis time, without any loss of separation or MS data quality. Downstream of the IMS separation drift tube, spatially dispersed ion packets are efficiently collected by another electrodynamic ion funnel and

transferred for analysis to an orthogonal acceleration TOF/MS analyzer stage. A high-performance data acquisition system based on a high speed analog-to-digital converter developed to ensure high mass accuracy, high dynamic range measurements is being used in conjunction with a real-time multi-dimensional spectral averaging capability developed under a new CRADA with Agilent Technologies.

Detailed evaluation of the LC-IMS-TOF/MS platform has confirmed significantly improved performance compared to the best currently available proteomics platforms. The new platform provides more than an order of magnitude increase in data generation rates, and initial studies confirm more than an order of magnitude improvement in sensitivity, as well as lower limits of detection. Further improvements in performance are expected from the use a more intense ion source based upon an advanced ESI multi-emitter design used in conjunction with a dual stage ion funnel interface. These advances are being complemented by the development of a new informatics pipeline for rapidly processing and analyzing the greatly expanded data volumes. In combination with improved informatics tools, application of the new platform is expected to enable much more comprehensive coverage of proteins (and e.g., modified proteins) and chemically diverse metabolites

This research is supported by the U.S. Department of Energy Office of Biological and Environmental Research at Pacific Northwest National Laboratory (operated for the U.S. Department of Energy by Battelle through Contract No. DE-AC05-76RLO 1830).

## 183 Stable Isotope Probing of RNA Combining Phylogenetic Microarrays and High Resolution Secondary Ion Mass Spectrometry to Link Composition and Function in Microbial Systems

Xavier Mayali<sup>1\*</sup> (mayali1@llnl.gov), Peter K. Weber,<sup>1</sup> Eoin L. Brodie,<sup>2</sup> Todd Z. DeSantis,<sup>2</sup> Ulas Karaoz,<sup>2</sup> Gary L. Andersen,<sup>2</sup> Meredith M. Blackwell,<sup>3</sup> Stephanie R. Gross,<sup>3</sup> Shalini Mabery,<sup>1</sup> Paul D. Hoeprich,<sup>1</sup> Ian D. Hutcheon,<sup>1</sup> and Jennifer Pett-Ridge<sup>1</sup>

<sup>1</sup>Lawrence Livermore National Laboratory, Livermore, Calif.; <sup>2</sup>Lawrence Berkeley National Laboratory, Berkeley, Calif.; and <sup>3</sup>Louisiana State University, Baton Rouge

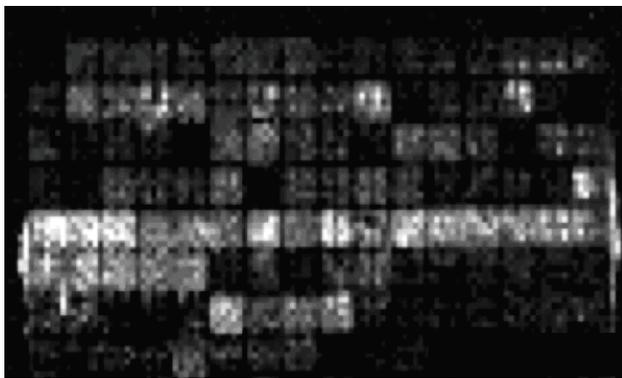
**Project Goals: To develop nano-scale stable isotope probing, complex community meta-transcriptomic analysis, and the translation of genome scale data into biogeochemical and metabolic flux network models.**

A fundamental goal in microbial ecology is to understand the biogeochemical role of individual microbial taxa in their natural habitat. This rather simple concept is in actuality a complex problem because 1) most microbes remain uncul-

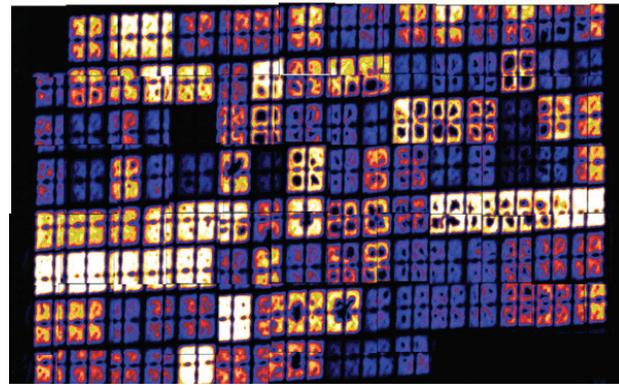
tivated and 2) the majority of microbial communities are very diverse. The former makes the direct testing of isolated strains for biogeochemical activity a limited approach. The latter impacts culture-independent methods like metagenomics as many biogeochemical processes cannot be directly inferred from sequence data alone, even when assembly of complete genomes is possible.

Our approach to this problem (Chip-SIP) involves the combination of high-density phylogenetic microarrays (“chips”) and stable isotope probing (SIP) to directly link identity and function. Microbial communities are incubated in the presence of stable isotope-enriched substrate (s), RNA is extracted and hybridized onto a microarray synthesized on a conductive surface, and the array is imaged using high resolution secondary ion mass spectrometry (SIMS) with a Cameca NanoSIMS 50 to detect isotopic enrichment. We have successfully validated this approach utilizing RNA from a single pure culture with varying degrees of isotopic enrichment using two different substrates ( $^{15}\text{N}$ -labeled ammonium and  $^{13}\text{C}$ -labeled glucose). We show that isotopic enrichment of individual probe spots as detected by nanoSIMS is positively correlated with fluorescence as detected by a traditional microarray scanner (figure 1). This allows the relationship between hybridization efficiency and relative isotopic enrichment to be determined. Further, we have successfully detected  $^{15}\text{N}$  enrichment in an estuarine bacterial community following incubation with  $^{15}\text{N}$ - $\text{NH}_4$ , demonstrating the utility of the method in mixed natural communities.

Current efforts are aimed at elucidating the major players in nitrogen fixation and carbon transformation in the gut of the wood-eating passalid beetle *Odontotaenius disjunctus*. The gut of this organism is spatially segregated into at least 4 distinct compartments (foregut, midgut, anterior hindgut, and posterior hindgut; see figure 2) each differing physically, chemically and microbiologically. We hypothesize that the sequential biogeochemical activities required to derive energy from lignocellulosic materials are partitioned across the gut sections. We are employing the Chip-SIP approach to determine which organisms at each gut location are involved in these processes.



Fluorescence by microarray scanner



$^{13}\text{C}$  enrichment by nanoSIMS

Figure 1: Visual comparison of  $^{13}\text{C}$ -enriched RNA from *Pseudomonas stutzeri* hybridized to an array comprised of *Pseudomonas*-specific probes, showing corresponding signal between fluorescence (top) and  $^{13}\text{C}$  isotopic enrichment by nanoSIMS (bottom)

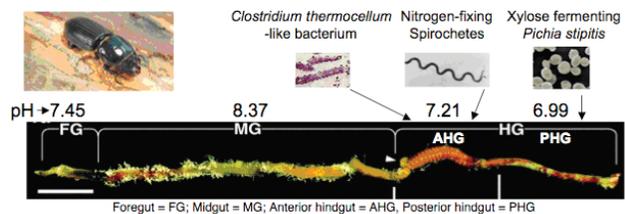


Figure 2: Dissected *Odontotaenius disjunctus* beetle gut showing the 4 sections (foregut, midgut, anterior hindgut and posterior hindgut); adapted from Nardi et al., Arthropod Struct. Dev. 35 (2006) 57-68. Scale-bar = 10 mm

## 184 Synthetic Genomics: Progress on Construction of a Synthetic Bacterial Cell

John Glass\* (jglass@jcvl.org)

Synthetic Biology Group, J. Craig Venter Institute, Rockville, Md.

**Project Goals:** Synthesize a minimal mycoplasma genome that has all of the machinery for independent life. Our goal in this aspect of the program is to create a minimal bacterial cell based on *Mycoplasma genitalium*, which has the smallest genome of any bacterial cell that can be grown in pure culture. A minimal cell contains only essential genes and can be grown in pure culture under defined conditions. It lacks synthetic capacity for small molecules or metabolites that can be supplied in the medium. Thus it is stripped down to core functions for macromolecular synthesis and cell division. The rationale for this is that through creation and analysis of a cell with perhaps fewer than 400 protein coding genes we will be better able to learn the first principals of cellular life. Such a cell would have less than one tenth as many genes as *Escherichia coli* and the lack of complexity would enable an uncluttered perspective on how cells work.

Bacteria and yeast have been widely used as hosts for cloning segments of DNA from a variety of organisms. Cloning of large DNA segments is limited by size and toxicity to the host. Reports of *Escherichia coli* DNA clones larger than three hundred kilobases have been infrequent, whereas yeast has been commonly used to clone megabase-sized DNA. We cloned whole bacterial genomes from *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, and *Mycoplasma mycoides* as circular centromeric plasmids in yeast. Once cloned, the bacterial genomes can be modified using efficient, well-established methods for DNA manipulation in yeast. Next, for one of those cloned genomes, *M. mycoides*, we introduced such modified genomes back into a different bacterial cell, *Mycoplasma capricolum*, by genome transplantation. To do this, restriction barriers had to be overcome. These methods should enable future transplantation of a synthetic genome, and also facilitate the engineering of bacteria with poorly developed genetic systems.

## 185

### Characterization of Microbial Strains Important in Biofuels and Biomass Conversion

Christopher L. Hemme,<sup>1\*</sup> Matthew W. Fields,<sup>2</sup> Qiang He,<sup>3,4</sup> Ye Deng,<sup>1</sup> Lu Lin,<sup>1,5</sup> Qichao Tu,<sup>1</sup> Housna Mouttaki,<sup>1</sup> Xueyang Feng,<sup>6</sup> Zheng Zuo,<sup>6</sup> B.D. Ramsey,<sup>2</sup> Zhili He,<sup>1</sup> Kerrie Barry,<sup>7</sup> Elizabeth Saunders,<sup>8</sup> Hui Sun,<sup>7</sup> Miriam Land,<sup>9</sup> Yun-Juan Chang,<sup>9</sup> Liyou Wu,<sup>1</sup> Joy Van Nostrand,<sup>1</sup> Loren Hauser,<sup>9</sup> Alla Lapidus,<sup>7</sup> Cliff S. Han,<sup>8</sup> Jian Xu,<sup>5</sup> Yinjie Tang,<sup>6</sup> Juergen Wiegand,<sup>10</sup> Tommy J. Phelps,<sup>11</sup> Eddy Rubin,<sup>7</sup> and Jizhong Zhou<sup>1,12</sup>

<sup>1</sup>Institute for Environmental Genomics, University of Oklahoma, Norman; <sup>2</sup>Dept. of Microbiology, Montana State University, Bozeman; <sup>3</sup>Dept. of Civil and Environmental Engineering and <sup>4</sup>Center for Environmental Technology, University of Tennessee, Knoxville; <sup>5</sup>Qingdao Institute of BioEnergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao, China; <sup>6</sup>Dept. of Energy, Environmental and Chemical Engineering, Washington University, St. Louis, Mo.; <sup>7</sup>DOE Joint Genome Institute, Walnut Creek, Calif.; <sup>8</sup>DOE Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, N.M.; <sup>9</sup>Genome Analysis and Systems Modeling Group, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; <sup>10</sup>Dept. of Microbiology, University of Georgia, Athens; <sup>11</sup>Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; <sup>12</sup>Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.

**Project Goals:** To explore the genomic and physiological properties of stable bacterial co-cultures important for biomass conversion and biofuels production in consolidated bioprocessing schemes.

#### Comparative Genomics of *Clostridia*

Genomic sequencing of 20+ *Clostridia* strains related to

biofuels production and biomass conversion were sequenced, including multiple strains from Cluster III thermophilic and mesophilic cellulolytic *Clostridium* species and multiple strains of saccharolytic *Thermoanaerobacter* species. This dataset represents a significant improvement in the genomic knowledge base of bacteria important to biofuels production. The genomes of two strains of *Thermoanaerobacter*, *T. pseudethanolicus* 39E and *Thermoanaerobacter* sp. X514, have been finished and comparative genomics analysis has been conducted. Experimental studies have shown that when either of these *Thermoanaerobacter* strains are grown in coculture with *Clostridium thermocellum* LQRI, cellulose degradation rates and ethanol production yields are increased compared to the *C. thermocellum* monoculture. Furthermore, this effect is greater for the X514 co-culture compared to the 39E co-culture. Comparative genomics and experimental analysis revealed several potential mechanisms by which such physiological effects could manifest. First, it was noted that X514 encodes a complete *de novo* Vitamin B<sub>12</sub> biosynthesis operon whereas 39E encodes only a partial operon. Experimental analysis showed that X514 monocultures are largely insensitive to the addition of exogenous B<sub>12</sub>, while ethanol yields are severely impacted in 39E monocultures when no exogenous B<sub>12</sub> is added. This effect can be alleviated in 39E when 2-3X B<sub>12</sub> is added to the culture. This effect is magnified in coculture with *C. thermocellum*, suggesting that X514 synthesizes its own B<sub>12</sub> while 39E does not and that B<sub>12</sub> is a critical nutrient in determining ethanol yields. Metabolic flux analysis also revealed that absolute flux through the central carbon metabolism pathway is greater in X514 than in 39E. Finally, 39E and X514 encode distinctly different xylose uptake systems and the X514 genes in general are more highly expressed under xylose growth conditions compared to 39E. Thus, substrate uptake, metabolic flux rates and vitamin synthesis likely contribute greater to variable ethanol production and cellulose degradation rates in *Thermoanaerobacter*-*C. thermocellum* co-cultures.

#### Hydrogen Production from *Desulfovibrio vulgaris*-*Clostridium cellulolyticum* Cocultures

Experimental analysis shows that the stable coculture of *D. vulgaris*-*C. cellulolyticum* produces significantly higher concentrations of molecular hydrogen when grown on cellulose compared to *C. cellulolyticum* monocultures. Functional genomic and experimental analyses were conducted to identify the mechanisms behind this observation. SEM images suggest that cellular binding to cellulose fibers is greater in the coculture than in the *C. cellulolyticum* monoculture, suggesting that the addition of *D. vulgaris* increases binding to the cellulose fibers and may in turn increase cellulose degradation rates. Preliminary microarray analysis also shows that *C. cellulolyticum* cellulosome genes are more highly expressed in co-culture than in monoculture as well as NiFe-hydrogenase genes and other genes related to hydrogen production.

#### Transcriptional Profiles of X514 at Different Carbon Substrates

The transcriptional profiles of *Thermoanaerobacter* sp. X514 at different carbon substrates have been conducted. Experimental studies show that X514 is able to metabolize

hexose (glucose, fructose, ribose, galactose and so on), pentose monosaccharides (including xylose) and some complex carbohydrates (sucrose, cellobiose, starch). When X514 are grown in glucose, xylose, fructose and cellobiose, the corresponding genes in carbon uptake system are more highly expressed. Moreover, X514 metabolized these four sugars by Embden-Meyerhof-Parnas (EMP) pathway and pentose phosphate (PPP) pathway. X514 encodes carbohydrate active enzymes for catabolism of fructose, xylose and cellobiose. In contrast to glucose metabolism, growth on fructose, xylose and cellobiose resulted in upregulation of carbohydrate metabolism genes which shift carbon fluxes head towards ribose. These observations suggest that when X514 is grown on fructose, xylose and cellobiose, more ribose should be synthesized as the substrate of nucleotide and amino acid metabolism. Experimental analysis shows that energy metabolism of X514 on fructose is more active than that on other sugars, with higher concentrations of ethanol, acetate and lactate generated. Furthermore, the V-type ATPase genes and a large number of genes involved in inorganic ion transport and metabolism (such as sodium-translocating decarboxylase enzyme genes, Na<sup>+</sup>/H<sup>+</sup> antiporter and sodium/hydrogen exchanger genes and so on) were significantly up-regulated. The data indicate that under fructose growth conditions, electrochemical ion gradient at the cytoplasmic membrane is much more actively established than when grown on other sugars. Thus, more ATP should be generated under these conditions. For the alcohol generation, the results show the three characterized *adh* genes are all expressed at similar levels when grown on these four sugars. But for the additional six lineage-specific *adh* genes, the expression levels greatly varied under different growth conditions, indicating differential expression of the *adh* genes in X514 under different growth conditions.

## 186

### Nitrate Reduction and Functional Characterization of *c*-type Cytochromes in *Shewanella*

Dongru Qiu,<sup>1\*</sup> Haichun Gao,<sup>1</sup> Zhili He,<sup>1</sup> Jingrong Chen,<sup>1</sup> Yili Liang,<sup>1</sup> Soumitra Barua,<sup>1,2</sup> Margaret F. Romine,<sup>3</sup> Samantha Reed,<sup>3</sup> Dave Culley,<sup>3</sup> David Kennedy,<sup>3</sup> Yunfeng Yang,<sup>2</sup> Kenneth H. Nealson,<sup>5</sup> James M. Tiedje,<sup>4</sup> James K. Fredrickson,<sup>3</sup> and Jizhong Zhou<sup>1,2</sup> (jzhou@ou.edu)

<sup>1</sup>Institute for Environmental Genomics and Dept. of Botany and Microbiology, University of Oklahoma, Norman; <sup>2</sup>Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; <sup>3</sup>Pacific Northwest National Laboratory, Richland, Wash.; <sup>4</sup>Center for Microbial Ecology, Michigan State University, East Lansing; and <sup>5</sup>Dept. of Earth Sciences, University of Southern California, Los Angeles

**Project Goals:** As part of the GTL program, our research is focused on nitrate reduction and functional

**characterization of *c*-type cytochromes in *Shewanella***  
*Shewanella* species are characterized by their respiratory versatility and psychrophility. Their ability to utilize a wide range of electron acceptors for respiration is due to a large number of *c*-type cytochromes in their genome. The dissimilatory metal reduction capacity of *Shewanella* and *Geobacter* provides a potential opportunity for the efficient bioremediation and electricity generation.

#### The NapC- and CymA-Dependent Nitrate Reduction in *Shewanella*

Nitrate respiration systems are highly diverse among *Shewanella* species. Bioinformatics analysis revealed three types of nitrate reduction systems in *Shewanella* genomes. *S. oneidensis* MR-1 harbors only the CymA-dependent nitrate reductase and the NapGH ubiquinol oxidase. Most *Shewanella* species, including *S. putrefaciens* W3-18-1, have both CymA- and NapC-dependent nitrate reductases, as well as the NapGH ubiquinol oxidase. The *S. baltica* strains have both the CymA- and NapC-dependent nitrate reductases but lack the NapGH ubiquinol oxidase. MR-1 appears to be atypical because it lacks both *napC* and *nrfBCD*, whose gene products act to transfer electrons from the quinol pool to terminal reductases NapA and NrfA. In *E. coli*, NapC and NrfBCD are essential for catalyzing reductions of nitrate to nitrite and the subsequent reduction of nitrite to ammonium, respectively. Our previous results revealed that CymA is likely to be a functional replacement of both NapC and NrfBCD in the nitrate and nitrite reduction in MR-1.

Our results suggest that the two-step manner of nitrate reduction found in MR-1 may be common among *Shewanella* species. Deletion of the *nap1* (*napDAGHB*) or *nap2* (*napDABC*) operon did not significantly affect cell growth, but the double mutant could not grow on nitrate, suggesting that the two *nap* operons are functionally redundant. In addition, the in-frame of *cymA* and *napC* of W3-18-1 deletion mutants did not show severe growth inhibition on nitrate, though deletion of *cymA* resulted in the loss of nitrate and nitrite reduction and growth in MR-1. Furthermore, the *cymA* deletion mutant showed little growth on nitrite in contrast to the *napC* deletion mutant, indicating that CymA was involved in nitrite reduction in both W3-18-1 and MR-1. The *cymA* gene from W3-18-1 complements the MR-1 *cymA* in-frame deletion mutant and allows reduction of ferric ions, nitrate, and nitrite when expressed *in trans*. The *napC* gene from W3-18-1 can also complement the MR-1 *cymA* deletion mutant and allows ferric iron reduction but it failed to allow nitrite reduction. These results support the hypothesis that the NapC-dependent and CymA-dependent periplasmic nitrate reduction systems allow an efficient dissimilatory reduction of nitrate and nitrite. Deletion of *narP* and *narQ* resulted in the growth inhibition on nitrate, suggesting that nitrate reduction is also regulated by the NarQP two-component system in W3-18-1. Our competition assays showed that W3-18-1 had a competitive advantage over MR-1 when grown together on nitrate.

### Characterization of *C*-type Cytochromes and Their Role in Anaerobic Respiration in *Shewanella*

The arsenal of *c*-type cytochromes is also highly diversified across the 21 sequenced *Shewanella* genomes and only twelve of the 41 *c*-type cytochrome of *S. oneidensis* MR-1 are present in all other sequenced strains. Only a few *c*-type cytochromes have been characterized. To discern the functions of unidentified *c*-type cytochrome genes in bacterial energy metabolisms, we generated 37 single mutants with an in-frame deletion of each individual cytochrome gene in MR-1. Reduction of a variety of electron acceptors was measured and the relative fitness was calculated for these mutants based on competition assays. This revealed that SO0610, SO1777, SO2361, SO2363, and SO4360 were important under aerobic growth conditions, and that most *c*-type cytochromes play a more important role in anaerobiosis. The *petC* gene appeared to be important to both aerobiosis and anaerobiosis. Our results regarding functions of CymA and MtrC are consistent with previous findings. We also assayed the biofilm formation of these mutants and results indicate that SO4666 might be important for pellicle formation.

*S. putrefaciens* W3-18-1 lacks orthologues for the secondary metal reductase and accessory proteins (MtrFED) of *S. oneidensis* MR-1. Sputw3181\_2446 encodes a decaheme *c*-cytochrome, orthologous to the outer membrane primary metal reductase OmcB of MR-1 (60% similarity) while another reductase similar to OmcA in MR-1 was also found in W3-18-1. Sputw3181\_2445 encodes an 11-heme *c*-type cytochrome OmcE, which only shares 40% similarity with the decaheme cytochrome OmcA. Single and double in-frame deletion mutants of *omcB* and *omcE* were generated for functional characterization of *omcE* and metal reduction in W3-18-1. Reduction of solid-phase Fe (III) and soluble Fe (III) in *S. putrefaciens* W3-18-1 was mainly dependent on OmcB under anaerobic conditions (with 50 mM lactate as electron donors and Fe<sub>2</sub>O<sub>3</sub>, α-FeO (OH), β-FeO (OH) and ferric citrate as electron acceptors. W3-18-1 catalyzed a more rapid reduction of α-FeO (OH) as compared to MR-1, suggesting that other genes may be involved in Fe (III) reduction in W3-18-1. As previously observed in MR-1, the deletion of both OmcE and OmcB led to a severe deficiency in reduction of solid-phase Fe (III) in W3-18-1 and an even greater deficiency in the reduction of soluble iron. The *omcB* and *omcE* genes of W3-18-1 have been expressed with the pBAD vector in *E. coli*. Heme staining assays also demonstrated that the disappearance of specific protein bands in the SDS-PAGE gels were consistent with *omcB* and *omcE* deletion in three mutant samples. These results suggest that *omcE* and *omcB* are actually expressed as cytochrome proteins and could play a central role in metal reduction in *S. putrefaciens* W3-18-1.

## 187

### Expression, Purification, and SAXS Structural Studies of the *Caulobacter* Chromosomal Segregation Machinery Complexes and Components

Jian Zhu,<sup>1</sup> Ping Hu,<sup>1</sup> Gary Anderson,<sup>1</sup> Thomas Earnest<sup>1\*</sup> (TEarnest@lbl.gov), and Harley McAdams<sup>2</sup>

<sup>1</sup>Lawrence Berkeley National Laboratory, Berkeley, Calif. and <sup>2</sup>Stanford University, Palo Alto, Calif.

**Project Goals: The purpose of this project is to: 1) map the protein-protein interaction network of *Caulobacter crescentus* including the implementation of TAP tagging to identify interaction partners for proteins that form biomolecular complexes; 2) isolate these complexes by co-expression in *E. coli*, overexpression in *Caulobacter* and/or cell-free systems and to automate these steps to achieve high-levels of throughput for complexes that will be applied on a pathway- to genome-scale; 3) characterize these complexes biochemically, functionally, and structurally at the molecular to cellular level.**

*Caulobacter crescentus* is stalked α-proteobacteria with significant demonstrated potential as a bioremediative agent, as well as being a well-characterized model organism for microbial systems biology. Approaches for the expression, purification, stabilization, and complex formation have been developed that allow for a high-success rate for the production of high-value target proteins and protein complexes. Chromosomal segregation during cell division is highly coordinated spatially and temporally through a set of proteins whose interactions drive the process of properly positioning the chromosomes to the two cells poles. In *Caulobacter* this utilizes the ParA and ParB proteins that interact with *parS* (proximal to the origin of replication). Additionally in *Caulobacter* these proteins interact with the novel cell polarity protein, PopZ, to attach the daughter chromosomes to the two cell poles through interactions of PopZ with ParB and *parS* (Bowman *et al.*, 2008). We express PopZ, ParA, and ParB individually in *E. coli* in sufficient quantities for structural and functional studies. SAXS studies on these proteins individually and in complex with each other reveals sets of interactions for these proteins. SAXS and native gel studies on PopZ alone indicates that in solution this ~20 kD proline-rich polypeptide forms a large assembly of ~ 300 kD which can be disassembled with urea and reassembled by urea removal *in vitro*. ParB forms a homo-dimer that becomes more compact with the addition of *parS*. ParA is an ATPase that oligomerizes upon ATP binding and interaction with ParB (Figure 1), and is thought to provide the force required to move the chromosomes to their proper position. The ParA/ParB/*parS* complex demonstrates a significant increase in the radius of gyration upon addition of ATP. The SAXS data have generated models for ParB, ParA and their complexes that can be compared and fitted with crystal structures of the components. These approaches, combined with the genetic, biochemical, and microscopic

data, are utilized to address structural and functional studies on a number of protein complexes involved in cell polarity, cell cycle control, transcriptional regulation, and bioremediation.

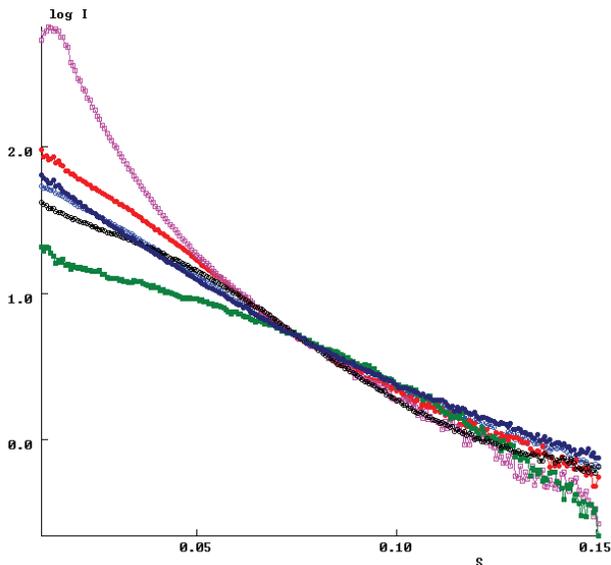


Figure 1. SAXS experiments on ParA, ParB, and *parS* demonstrate oligomerization upon addition of ATP (purple curve) compared with the sample without ATP (red curve). We thank Greg Hura of the ALS SIBYLS beamline 12.3.1 for assistance with the SAXS experiments.

## Biological Systems Interactions

# 188

### PNNL Foundational Scientific Focus Area— Biological Systems Interactions

**Jim Fredrickson**<sup>1\*</sup> (jim.fredrickson@pnl.gov), **Gordon Anderson**<sup>1</sup>, **Scott Baker**<sup>1</sup>, **Alex Beliaev**<sup>1</sup>, **Mary Lipton**<sup>1</sup>, **Jon Magnuson**<sup>1</sup>, **Margie Romine**<sup>1</sup>, **Thomas Squier**<sup>1</sup>, **H. Steven Wiley**<sup>1</sup>, Don Bryant,<sup>2</sup> Frank Collart,<sup>3</sup> William Inskeep,<sup>4</sup> Francois Lutzoni,<sup>5</sup> Andrei Osterman,<sup>6</sup> Margarethe Serres,<sup>7</sup> Harvey Bolton Jr.,<sup>1</sup> and David Ward<sup>3</sup>

<sup>1</sup>Pacific Northwest National Laboratory, Richland, Wash.; <sup>2</sup>Pennsylvania State University, University Park; <sup>3</sup>Argonne National Laboratory, Argonne, Ill.; <sup>4</sup>Montana State University, Bozeman; <sup>5</sup>Duke University, Durham, N.C.; <sup>6</sup>Burnham Institute for Medical Research, La Jolla, Calif.; and <sup>7</sup>Marine Biological Laboratory, Woods Hole, Mass.

**Project Goals: The primary FSFA objectives include—develop a mechanistic understanding of metabolic interactions among key members of microbial mats using**

**the tools of genomics and systems biology; understand the collective energy, carbon, and nutrient processing in laboratory-based microbial systems that contributes to their stability and efficient utilization of resources using a systematic application of -omics (transcriptomic, proteomic, and metabolomic) approaches; determine interspecies co-adaptations and functional innovations that contribute to robustness and functional efficiency; explore microbe-microbe and microbe-environment interactions that control genome evolution; determine the functional content of the mobile pool of genes in microbial mats and corresponding mechanisms by which they are disseminated; and understand cellular strategies that permit a system of interacting organisms to control the excess generation of reactive oxygen species to promote adaptive responses that enhance their survival; and systems biology investigations of the culturable lichen *Cladonia grayi* to understand mechanisms of resilience against environmental stress.**

The PNNL Genomic Science Foundational Scientific Focus Area (FSFA), initiated at the beginning of FY10, is addressing critical scientific issues on microbial interactions, investigating how microorganisms interact to carry out, in a coordinated manner, complex biogeochemical processes such as the capture and transfer of light and chemical energy. The primary research emphasis will be on associations between autotrophic and heterotrophic microorganisms with the additional objective of obtaining a predictive understanding of how interactions impart stability and resistance to stress, environmental fitness, and functional efficiency. The main scientific objectives of the FSFA include: development of a mechanistic understanding of interactions among key members of microbial autotroph-heterotroph associations (AHA) using the tools of genomics and systems biology; understanding the collective energy, carbon, and nutrient processing in AHAs that contributes to their stability and efficient utilization of resources; probing interspecies co-adaptations and functional innovations that contribute to robustness and functional efficiency and exploring the types of microbe-microbe and microbe-environment interactions that control genome evolution; understanding cellular strategies that permit a system of interacting organisms to control the excess generation of ROS to promote adaptive responses that enhance their survival; and systems biology investigations of the culturable lichen *Cladonia grayi* to understand mechanisms of resilience against environmental stress.

Autotroph-heterotroph microbial associations formed the foundation of the biosphere nearly 3 billion years ago with oxygenic photosynthetic prokaryotes (cyanobacteria) and their associated heterotrophic partners colonizing shallow ocean zones. The photolithotrophs use sunlight for energy to fix CO<sub>2</sub> and N<sub>2</sub> and produce O<sub>2</sub>, H<sub>2</sub>, and organic molecules that supported the growth and metabolism of their heterotrophic partners that facilitate recycling of carbon and nutrients. Autotroph-heterotroph associations are common planet-wide, representing metabolically interactive, self-sustaining communities that are often pioneering and can represent the only biota in extreme environments. These associations are well-adapted to a range of harsh conditions

that include extremes of temperature, salinity, desiccation, irradiance, high O<sub>2</sub>, and nutrient deprivation. Microbial associations, inclusive of photolithotrophs (e.g., light energy) and chemolithotrophs (e.g., inorganic chemical energy), are highly relevant to DOE mission areas including bioenergy, carbon cycling/sequestration, and contaminant fate and transport. Further, interacting microorganisms provide key services such as carbon, nutrient, and metal cycling to the biosphere, have considerable potential for a wide range of biotechnological applications, and present challenging and exciting new basic research opportunities.

The PNNL FSFA is utilizing genome-enabled systems biology approaches on three levels—molecular, cellular, and community—to elucidate the underlying design principles of microbial associations, emphasizing interactions between microorganisms for which there are established or hypothesized interdependencies. As part of this approach, the FSFA is developing the experimental tools and data necessary to quantitatively understand and predict causal relationships between environmental change, microbial associations, and cellular functions. The FSFA is using a combined top-down/bottom-up approach where bioinformatics-based genome functional predictions are made using a range of tools and resources; evaluations of evolutionary and ecological adaptation processes are made at the genome scale, high-throughput expression analyses and functional genomics are used to uncover key genes and proteins as well as metabolic and regulatory networks. The bottom-up component uses genetic, physiological, and biochemical approaches to test or verify predictions made by the top-down approaches. The top-down experimental component includes the generation of large amounts of data from biological perturbation experiments that support computational analyses to develop models of various cellular networks. The FSFA is utilizing a series of lab-based model systems consisting of constructed consortia with engineering potential, natural communities, and consortia derived from natural communities for hypothesis testing. Natural communities include microbial mats, biofilms, and lichens. These systems include associations that have evolved to permit successful colonization of extreme environments through effective utilization of solar and chemical energy and scarce nutrients. Research involving these systems will guide our ability to understand and predict how biological associations function with a high degree of efficiency and resiliency. A significant advantage is afforded by using a combination of mechanistic and systems-level investigations of representative associations cultivated in the laboratory under controlled conditions and analyses of natural assemblages using analytical and computational tools of systems biology.

# 189

## Transcriptional Regulation of *Shewanella* Central Carbon Metabolism by HexR

Dmitry Rodionov<sup>1,4\*</sup> (rodionov@burnham.org), Xiaoping Li,<sup>1</sup> Samantha Reed,<sup>2</sup> Margaret Romine,<sup>2</sup> James Fredrickson,<sup>2</sup> Pavel Novichkov,<sup>3</sup> Semen Leyn,<sup>4</sup> and Andrei Osterman<sup>1,5</sup> (osterman@burnham.org)

<sup>1</sup>Burnham Institute for Medical Research, La Jolla, Calif.; <sup>2</sup>Pacific Northwest National Laboratory, Richland, Wash.; <sup>3</sup>Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.; <sup>4</sup>Institute for Information Transmission Problems RAS, Moscow, Russia; and <sup>5</sup>Fellowship for Interpretation of Genomes, Burr Ridge, Ill.

**Project Goals:** This project was started as a component of the *Shewanella* Federation studies aimed at better understanding of the ecophysiology and speciation of this important genus. It is now continued in the framework of the PNNL Foundational Scientific Focus Area (FSFA) “Biological Systems Interactions” (PI, J. Fredrickson). This FSFA has a scientific focus on understanding interactions between microbes using systems biology approaches including state-of-science technologies, and it is focused on interactions between microbes and their extracellular environments, with an emphasis on acquiring an understanding of microbial autotroph-heterotroph associations. The predictive understanding of these biological systems will be acquired by integrating experimental and computational approaches that exploit the expertise from multiple disciplines working in a synergistic manner. Genomics and proteomics approaches will be applied to predict gene occurrence and function and to identify, quantify, and characterize individual proteins and complexes. Microbial physiology, ecology, biochemistry, metabolic profiling and genetic approaches will be used to verify predictions and test specific hypotheses.

The absence of genes encoding phosphofructokinase in all sequenced *Shewanella* leads to the Entner-Doudoroff (ED) and pentose phosphate (PP) pathways being the major routes of sugar utilization rather than a glycolytic route for central carbohydrate metabolism (CCM). Not surprisingly, such a redistribution of catabolic flux is associated with a completely different regulatory strategy compared to those used for classical *glycolytic* metabolisms found in other proteobacteria such as *Escherichia coli*. Using a comparative genomics approach, we have identified a novel *Shewanella* regulon that is controlled by HexR and that encompasses ~30 genes from the CCM pathways, as well as the deoxy-nucleoside and glycine utilization. The HexR-binding motif was predicted to be a 17-bp palindromic sequence with the consensus tTGTAATwwwATTACa. Assay of purified HexR protein by electrophoretic mobility shift analysis confirmed recognition of the predicted binding motifs by this regulator. The ED pathway intermediate, 2-keto-3-deoxy-6-phosphogluconate, functions as a HexR antagonist releasing it from its target operator. Analysis of the relative

position of the HexR binding sites and candidate promoters in multiple *Shewanella* genomes suggested a dual mode of HexR action; negative regulation (repression) of some of the target genes and positive regulation (activation) of others. This observation is in agreement with the expression patterns of 27 predicted HexR regulon genes observed in the ~200 *S. oneidensis* MR-1 microarray experiments available in the M3D database (<http://m3d.bu.edu/>).

Overall, three distinct groups of highly correlated HexR-regulated genes were revealed: (i) *zwf-pgl-edd-eda, pykA, tal-pgi, gapA2*; (ii) *phk, deoAB, cdd, nqrABCDE*; and (iii) *ppsA, gapA3, gcvTHP*. Remarkably, the third group of genes showed a strong anti-correlation with the first two groups supporting the proposed dual mode of HexR regulation. This observation was directly supported by qPCR-based comparison of the expression of HexR regulon genes in the wild-type and a targeted *hexR* deletion mutant of *S. oneidensis*. The most significant differences in WT vs. mutant gene expression patterns were observed between genes involved in catabolic pathways and in gluconeogenesis (repressed or activated, respectively). For example, of the two genes, *pykA* and *ppsA*, that encode enzymes catalyzing phosphoenolpyruvate to pyruvate interconversion in opposite directions, the former is repressed whereas the latter is activated by HexR. Comparison of growth phenotypes of mutant and wild type strains on various carbon sources (N-acetylglucosamine, glycerate, inosine, and lactate) showed that *hexR* deletion leads to an inability of *S. oneidensis* to grow on lactate as a single carbon source. This finding confirmed the observed positive mode of action of the HexR regulator on the gluconeogenic gene, *ppsA*, whose activity is known to be essential for the growth of *E. coli* on lactate. The detailed results of our HexR regulon reconstruction, including the predicted transcription factor binding sites, are presented in a recently developed RegPrecise database (<http://regprecise.lbl.gov>).

Additional physiological studies and metabolomic profiling analyses are in progress to further investigate the role of HexR in the regulation of CCM in *Shewanella*. The HexR regulon in *Shewanella* may be considered as a partial functional replacement of a classical 6-fructose-phosphate regulon FruR, which is known to control fructose utilization and CCM in *E. coli*. The sequenced *Shewanellae* lack FruR and are not able to grow on fructose. Reconstruction and comparative analysis of HexR regulons was expanded to a broader set of genomes from  $\gamma$ - and  $\beta$ -proteobacteria and some Firmicutes contributing to better understanding of evolutionary history of HexR and its role in the regulation of CCM.

These studies demonstrate the value of applying comparative genomics and complementary experimental analyses to predict and validate regulatory networks in previously uncharacterized biological systems. As part of the new Foundational Science Focus Area project team led by the Pacific Northwest National Laboratory we intend to continue applying such strategies to explore regulatory networks in individual species and communities.

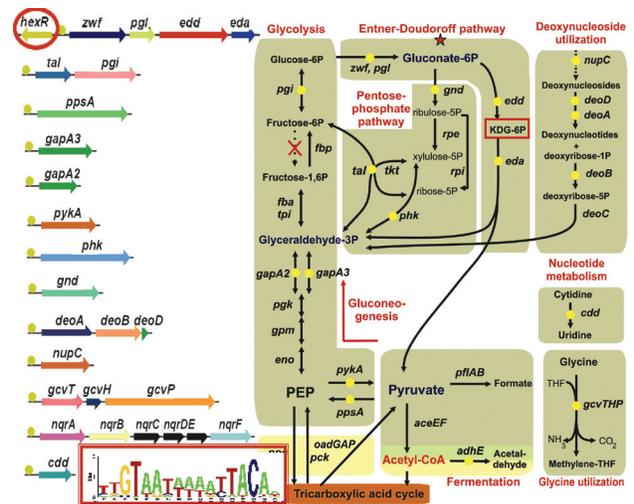


Fig. 1. The reconstructed HexR regulon and target pathways in *Shewanella oneidensis* MR-1.

# 190

## Finding Function for Fungal Glycoside Hydrolases

Jon K. Magnuson,<sup>1</sup> Frank Collart,<sup>2</sup> Deanna Auberry,<sup>1</sup> Sarah Zerbs,<sup>2</sup> Ellen A. Panisko,<sup>1</sup> Justin Powlowski,<sup>3</sup> Adrian Tsang,<sup>3</sup> and **Scott E. Baker**<sup>1\*</sup> ([scott.baker@pnl.gov](mailto:scott.baker@pnl.gov))

<sup>1</sup>Pacific Northwest National Laboratory, Richland, Wash.; <sup>2</sup>Argonne National Laboratory, Argonne, Ill.; <sup>3</sup>Concordia University, Montreal, Quebec, Canada

**Project Goals: The goal of this project is to develop a pipeline for functional annotation of fungal glycoside hydrolases.**

Plant biomass is efficiently broken down and used a source of carbon by many microbes, including fungi. Many enzymes that are involved in the deconstruction of cellulose and hemicelluloses are members of a large group of enzymes called glycoside hydrolases (GHs). On average, the genomes of filamentous fungi contain well over 100 genes encoding different GH family enzymes. Currently, GHs are classified into families based on sequence and predicted structure. GH families may contain multiple enzymatic activities. The goal of our project is to develop a pipeline for functional characterization of GHs. Functional information GHs is generated by genome annotation, proteomic analysis of the secretome and enzymatic assays of expressed GHs of interest. We have cultured on different substrates and subsequently processed a variety of fungal secretomes for proteomic analysis. We compare protein expression profiles of GHs across different culture substrates. Finally, we have compared the activity and stability of expressed *Aspergillus niger* enzymes produced by different hosts: *E. coli*, *Pichia* and *Aspergillus niger*. Our data allow us to add functional data to annotations of fungal

GHs there were previously characterized based solely on predicted amino acid sequence analysis.

## 191

### High-Temperature Chemotrophic Microbial Communities of Yellowstone National Park: Metagenomics Provides a Foundation for Dissecting Microbial Community Structure and Function

**W. Inskeep**<sup>1\*</sup> (binskeep@montana.edu), Z. Jay, M. Kozubal,<sup>1</sup> J. Beam,<sup>1</sup> R. Jennings,<sup>1</sup> H. Bernstein,<sup>2</sup> R. Carlson,<sup>2</sup> D. Rusch,<sup>3</sup> and S. Tringe<sup>4</sup>

<sup>1</sup>Dept. of Land Resources and Environmental Sciences and Thermal Biology Institute and <sup>2</sup>Dept. of Chemical and Biological Engineering, Montana State University, Bozeman; <sup>3</sup>J. Craig Venter Institute, Rockville, Md.; and <sup>4</sup>DOE Joint Genome Institute, Walnut Creek, Calif.

**Project Goals:** The goals of this project are to study microbial interactions in model geothermal microbial communities. Extreme geochemical conditions and high temperature result in low-diversity microbial communities where metagenomic sequence data can be used to dissect microbial community structure and function.

Microbial communities are a collection of interacting populations. However, a significant fraction of our knowledge base in microbiology originates from organisms grown and studied in pure culture, in the absence of other members of the community who may compete for resources or provide necessary co-factors and or substrates. Moreover, many of the organisms studied in pure culture have not necessarily represented the numerically dominant members of microbial communities found in situ. The advent of molecular tools (e.g., genome sequencing) has provided opportunities for assessing the predominant and relevant indigenous organisms, as well as their likely function within a connected network of different populations (i.e., community). High-temperature microbial communities are often considerably less diverse than mesophilic environments and constrained by dominant geochemical attributes such as pH, dissolved oxygen, Fe, sulfide, and or trace elements including arsenic and mercury. Consequently, the broader goal of our work is to utilize extreme high-temperature geothermal environments including acidic Fe-oxidizing communities as model systems for understanding microbial interactions among community members. Recent metagenomic sequencing of high-temperature, acidic Fe-mats of Norris Geyser Basin, Yellowstone National Park (YNP), conducted as part of a DOE-Joint Genome Institute (Community Sequencing Project) reveal communities dominated by novel members of the *Archaea*, bacterial members of the deeply-rooted Order Aquificales as well as other less-dominant Bacillales and Clostridiales. Phylogenetic and functional analysis of metagenome sequence is providing an excellent foundation for establishing the role of individual populations in a net-

work of interacting community members, and for directing hypotheses regarding the importance of specific biochemical pathways responsible for material and or energy cycling. For example, we are using metagenome sequence in combination with information available from reference strains to identify protein-coding sequence of importance in the oxidation and or reduction of Fe, S, O, and As, as well as central C metabolism (including fixation of CO<sub>2</sub>). Genes coding for proteins with hypothetical or putative roles in electron transfer, C-capture and C-transformation have been prioritized for design of quantitative-reverse transcriptase-PCR (Q-RT-PCR) primers to evaluate functional capacity quantitatively in both pure-culture and subsequent mixed communities. Future proteomic and transcriptomic analyses, as an element within the PNNL Foundational Scientific Focus Area, will focus on both pure-culture experiments under different electron donor and acceptor conditions, as well as natural thermophilic mats. Proteomic results will be used to assess and confirm the importance of specific proteins and to improve microbial community models. Application of genomic, proteomic, and metabolic information to dissect microbial community structure and function is tractable within high-temperature geothermal systems in part due to the relative simplicity of the community and the dominance of several key geochemical variables (i.e. pH, Fe, O<sub>2</sub>).

## 192

### Comparative Genome Analyses of Members of the Ecologically Versatile Genus *Shewanella*: Searching for Sequence Signatures That Reflect Environmental Adaptation

**Margrethe H. Serres**<sup>1\*</sup> (mserres@lbl.gov), Raghu P.R. Metpally,<sup>1,2</sup> and Margaret F. Romine<sup>3</sup>

<sup>1</sup>Marine Biological Laboratory, Woods Hole, Mass.; <sup>2</sup>University of Iowa, Iowa City; and <sup>3</sup>Pacific Northwest National Laboratory, Richland, Wash.

**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.

*Shewanellae* are an environmentally important group of bacteria whose members have been isolated from many different ecosystems (fresh and marine waters and sediments, a deep sea iron mat, subsurface sedimentary rock, and squid gland) that vary in atmospheric pressure, temperature, and salinity. These organisms thrive in red-ox interfaces in the environment and are well known for their versatile respiratory capability, using over 20 different compounds as electron acceptors. Complete genome sequences have been obtained for 19 *Shewanellae*. The strains were selected for sequencing based on their phylogenetic distance, some representing closely related sub-species clusters and others being more distantly related. Overall they represent a genetic gradient in which speciation and environmental adaptation

can be studied. A recent comparative analysis of the genome sequences, phenotypic characteristics, and proteomic expression profiles of the first ten strains sequenced showed that phenotypic and genotypic similarities largely correlated with phylogenetic distances despite the evidence of laterally transferred genes (1). Some of the phenotypic and genotypic traits were more conserved with increasing evolutionary distance (i.e. predicted metabolic pathways) than others (i.e. protein expression patterns).

Our comparative analysis has extended to 19 completed *Shewanella* genome sequences, including more distantly related strains that are obligately marine or that thrive in lower or higher temperatures than those previously studied. The protein sequences from all 19 strains have been analyzed for their domain content (Pfam, TIGRFam) in order to detect differences (functional, protein family compositions) that correlates with the environment in which the strains were isolated. Protein families involved in responses to environmental factors (chemotaxis proteins, two-component regulators, signaling proteins) appear to be large and more diverse among the *Shewanellas*. A curated table of orthologous proteins has been generated for the sequenced shewanellae allowing for categorization of *Shewanella* proteins as core (present in all strains), dispensable (absent in one or more strains), or strain-specific. A significant amount of curation relating to gene calling and function assignments has been done for this dataset, which is available in the *Shewanella* Knowledgebase (2). Of the 16612 orthologous groups of *Shewanella* genes (redundant genes removed), 11% are core genes, 54% dispensable genes and 35% unique genes, the two latter categories encoding genes involved in adaptation to and survival in select environments.

In addition to studying the presence and absence of genes, we are now searching the *Shewanella* genomes for evidence of selective changes in the sequences that can be linked to our knowledge of specific functions or specific environmental conditions of the strains. Sequence changes at non-synonymous vs. synonymous sites have been identified to find genes that are under a purifying or a diversifying selection pressure. Amino acid replacement ratios, radical vs. conservative changes, have also been determined (3). Such codon usage analyses have been used to identify genes that are under biochemical or ecological constraints. We have also calculated the Codon Adaptation Indexes for the *Shewanella* sequences, an estimate of the synonymous codon usage bias and of gene expression levels. The *Shewanella* ortholog table is used as a framework for our studies allowing us to separate the genes into sets that are common to all of the *Shewanellae* or that vary among the strains giving them their unique characteristics. The sequence changes are also evaluated relative to the gene product functions, locations, and protein family memberships.

Our comparative analysis of members of the *Shewanella* genus is forming the foundation for studying other groups of related organisms as well as consortia of microbes in selected environments.

## References

1. Konstantinidis KT, Serres MH, Romine MF, Rodrigues JL, Auchtung J, McCue LA, Lipton MS, Obraztsova A, Giometti CS, Neelson KH, Fredrickson JK, Tiedje JM. 2009. Comparative systems biology across an evolutionary gradient within the *Shewanella* genus. Proc Natl Acad Sci U S A. 106 (37):15909-14.
2. www.shewanella-knowledgebase.org
3. Hanada K, Shiu SH, Li WH. 2007. The nonsynonymous/synonymous substitution rate ratio versus the radical/conservative replacement rate ratio in the evolution of mammalian genes. Mol Biol Evol. 10:2235-41.

# 193

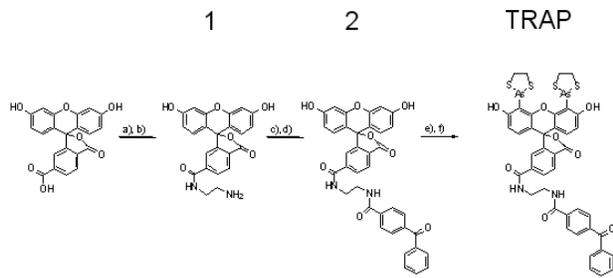
## In-Vivo Trapping and Structural Interrogation of Transient Protein Complexes

M. Uljana Mayer, Ping Yan, Ting Wang, Yijia Xiong, Diana J. Bigelow, and **Thomas C. Squier\*** (thomas.squier@pnl.gov)

Pacific Northwest National Laboratory, Richland, Wash.

**Project Goals: Identification of new imaging methods that permit high-throughput measurements of protein complexes that will allow the development of predictive models for bioenergy applications and with regard to how microbes respond to environmental change.**

Protein cross-linking, especially coupled to mass-spectrometric identification, is increasingly used to determine protein binding partners and protein-protein interfaces for isolated protein complexes. The modification of cross-linkers to permit their targeted use in living cells is of considerable importance for studying protein-interaction networks, which are commonly modulated through weak interactions that are formed transiently to permit rapid cellular response to environmental changes. We have therefore synthesized a targeted and releasable affinity probe (TRAP) consisting of a biarsenical fluorescein linked to benzophenone that binds to a tetracysteine sequence in a protein engineered for specific labeling (Scheme 1). Here, the utility of TRAP for capturing protein binding partners upon photoactivation of the benzophenone moiety has been demonstrated in living bacteria and eukaryotic cells. In addition, ligand exchange of the arsenic-sulfur bonds between TRAP and the tetracysteine sequence to added dithiols results in fluorophore transfer to the crosslinked binding partner. Following isolation of protein complexes, the facile release of TRAP from the original binding site permits the identification of the proximal binding interface through mass spectrometric fragmentation and computational sequence identification.



Scheme 1. *Synthesis of TRAP.* A) EDC, Et<sub>3</sub>N, NHS, dry DMF, 30 min; B) *N*-Boc-ethylenediamine, 16 h; C) 20 % TFA/CH<sub>2</sub>Cl<sub>2</sub>, 2 h; D) 4-Benzoylbenzoic acid, EDC, NHS, iPr<sub>2</sub>EtN, DMF, 16 h; E) HgO, TFA, 70 °C; F) 1) AsCl<sub>3</sub>, PdOAc, iPr<sub>2</sub>EtN, NMP, 4 h; 2) EDT, 20 % acetone/H<sub>2</sub>O (overall yield: 2 %).

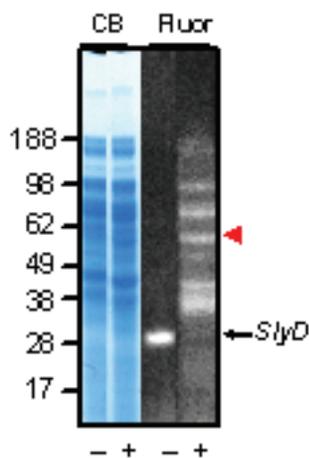


Figure 1: *Selective in-vivo labeling of chaperon SlyD and Photo-dependent Trapping of Binding Partners that Include Elements of [Ni-Fe]-Hydrogenase Maturation Pathway.* Coomassie Blue (CB) protein stain (left panel) and fluorescence image (right panel) before (-) and following (+) photodependent cross-linking. Positions of SlyD and HypB (red arrow), previously shown to bind SlyD, are indicated.

One example of the utility of using TRAP to identify molecular pathways associated with the maturation of [Ni-Fe]-hydrogenases involves the introduction of a tetracycysteine tag into the chaperone SlyD expressed in *Shewanella*, permitting its selective labeling (Figure 1). Following photoactivation and reduction, the fluorescence TRAP moiety is readily transferred to binding partners – resulting in a considerable simplification in the identification of the unique molecular interfaces following proteolytic digestion and mass spectrometric analysis (Figure 2).

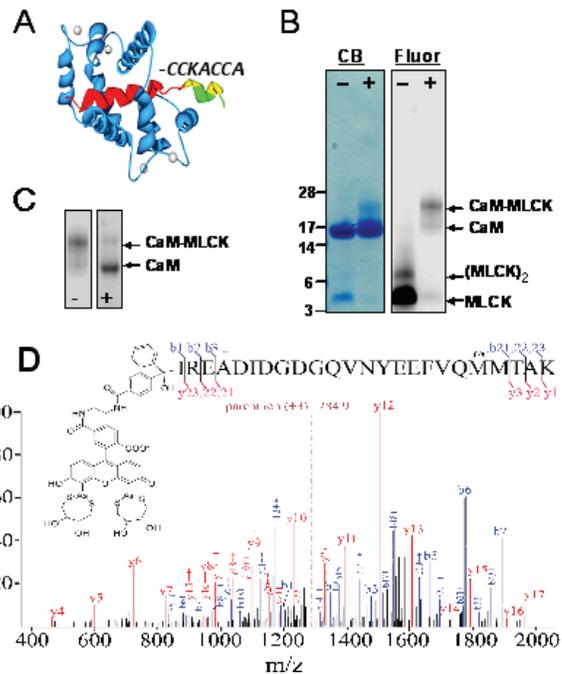


Figure 2: *Mass Spectrometric Identification of Interfacial Binding Site Following Crosslinking and Fluorophore Transfer.* (A) Depiction of the structural interface between tagged MLCK protein (red) containing engineered TRAP binding site CCKACCA (yellow and green) in complex with binding partner CaM (blue). (B) Coomassie Blue (CB) protein stain and fluorescence from TRAP reagent (Fluor) prior to (-) and following (+) photo-induced crosslinking. (C) Cross-linked complex before (-) and following (+) transfer of TRAP to binding partner upon addition of DTT (5 mM). (D) Mass spectrum of TRAP-bound peptide, where all b (blue) and y (red) fragment ions are identified. Site of TRAP binding in CaM bait protein is shown above mass spectrum.

**Conclusions:** We have synthesized a new class of photocrosslinker (i.e., TRAP) that can be targeted to a tetracycysteine tag to capture (that is, trap) protein binding partners upon light activation following the straight-forward introduction of a unique tetracycysteine binding sequence onto a protein of interest. The simplicity of this method will facilitate the high-throughput identification of protein complexes in a range of different organisms. The small size of TRAP compared to other targeted and multifunctional crosslinkers enables facile identification of the site of cross-linking by searching for the added molecular weight of the transferred crosslinker using commonly available mass spectrometers and publically available software. Optimal conditions associated with cross-linking complexes of interest are facilitated by the high-fluorescence yield of the TRAP reagent, enabling facile visualization of TRAP fluorescence on SDS-PAGE gels.

**Future Measurements:** TRAP will be used in combination with targeted in vivo photocrosslinking and mass spectrometry to identify interfacial binding sites, permitting an understanding of how environmental conditions affect protein-protein interaction networks. Coupled with the use of complementary multiuse affinity reagents (MAPs)

that permit the visualization of protein localization within microbial communities, we propose to use these reagents to identify regulatory elements that modulate energy flux through key metabolic pathways associated with biomass and the generation of biofuels.

### Reference

1. Yan P, T. Wang, G.J. Newton, T.V. Knyushko, Y. Xiong, D.J. Bigelow, T.C. Squier, and M.U. Mayer (2009) *A targeted releasable affinity probe (TRAP) for in vivo photocrosslinking*. *Chembiochem*. 10: 1507-1518.

## Plant-Microbe Interfaces

# 194

## Plant-Microbe Interfaces

**Mitchel J. Doktycz**<sup>1\*</sup> (doktyczmj@ornl.gov), Gerald A. Tuskan,<sup>2</sup> Christopher W. Schadt,<sup>1</sup> Gregory B. Hurst,<sup>3</sup> Edward Uberbacher,<sup>1</sup> Dale A. Pelletier,<sup>1</sup> Jennifer Morrell-Falvey,<sup>1</sup> Timothy J. Tschaplinski,<sup>2</sup> David J. Weston,<sup>2</sup> Scott T. Retterer,<sup>1</sup> Andrey Gorin,<sup>4</sup> Yunfeng Yang,<sup>1</sup> Robert Hettich,<sup>3</sup> Udaya C. Kalluri,<sup>2</sup> Xiaohan Yang,<sup>2</sup> Abhijit Karve,<sup>2</sup> Mircea Podar,<sup>1</sup> Steven D. Brown,<sup>1</sup> Robert Cottingham,<sup>1</sup> Tatiana Karpinets,<sup>1</sup> Chongle Pan,<sup>4</sup> Guru Kora,<sup>4</sup> Denise Schmoyer,<sup>1</sup> and Susan Holladay<sup>1</sup>

<sup>1</sup>Biosciences Division, <sup>2</sup>Environmental Sciences Division, <sup>3</sup>Chemical Sciences Division, and <sup>4</sup>Computer Science and Mathematics, Oak Ridge National Laboratory, Oak Ridge, Tenn.

<http://PML.ornl.gov>

**Project Goals (Abstracts 194-205): Understand the genome-dependent molecular and cellular events involved in establishing and maintaining beneficial interactions between plants and microbes. *Populus* and its associated microbial community serve as an initial test system for understanding how these molecular events manifest themselves within the spatially, structurally, and temporally complex scales of natural systems. To achieve this goal, we will focus on 1) characterizing the natural variation in *Populus* microbial communities within complex environments, 2) elucidating *Populus*-microbial interactions at the molecular level and dissecting the signals and pathways responsible for initiating and maintaining microbial relationships, and 3) performing metabolic and genomic modeling of these interactions to aid in interpreting the molecular mechanisms shaping the *Populus*-microbial interface.**

Rapid progress in biological and environmental sciences has been enabled by the availability of genome sequences and the tools and technologies involved in interpreting genome function. As our understanding of biological systems grows,

it becomes increasingly clear that the functional expression of individual genomes is affected by an organism's environment and the community of organisms with which it associates. The beneficial association between plants and microbes exemplifies a complex, multi-organism system that is shaped by the participating organisms and the environmental forces acting upon it. These plant-microbe interactions can benefit plant health and biomass production by affecting nutrient uptake, influencing hormone signaling, effecting water and element cycling in the rhizosphere, or conferring resistance to pathogens. Studying the integral plant-microbe system in native, perennial plant environments, such as *Populus* and its associated microbial community, provides the greatest opportunity for discovering plant-microbial system functions relevant to DOE missions related to bioenergy and carbon-cycle research and understanding of ecosystem processes.

The functional attributes of *Populus* depend on the microbial communities with which it associates. Bacteria and fungi can be found within *Populus* tissues and closely associated with the roots in the rhizosphere. Understanding these communities, and the interfaces between organisms, is critical to realizing fundamental scientific knowledge that may enable increased plant productivity, ecosystem sustainability, disease resistance, drought tolerance, and ecosystem carbon budgets. This interface can also influence the processes, or mechanisms, by which adaptive traits arise from genetic variation and community function. Microbial rhizosphere structure, plant root bacterial and fungal colonization patterns, and the microbe-plant signaling pathways inherent in each type of association are all found within *Populus* and can be functionally translated hierarchically across scales into ecosystem patterns and processes.

Understanding the mechanisms by which plants and microbes interact represents a grand challenge facing biological and environmental science. How microbial selection and colonization occurs, what reciprocal benefits are bestowed upon the plant and microbe, and how these interactions ultimately affect, and are affected by, the environment are just some of the intrinsic scientific questions. The multiple spatial and temporal scales involved in these interfaces, the complexity of the component systems, and the need for better tools that use and build upon growing genomics resources to probe and interpret these combined systems represent some of the essential technical challenges. The variety and magnitude of these challenges are only offset by the impact and benefit of overcoming these challenges and in applying this understanding to issues as diverse as efficient energy transformation and carbon cycling.

195

Student Presentation

### Plant-Microbe Interfaces: Characterization of Cell Surface Properties in *Azospirillum brasilense* Wildtype Cells and Che1 Pathway Mutants Using Atomic Force Microscopy

A. Nicole Edwards,<sup>1\*</sup> Piro Siuti,<sup>1</sup> Jennifer L. Morrell-Falvey,<sup>1,2</sup> (morrellj1@ornl.gov), Amber N. Bible,<sup>3</sup> Gladys Alexandre,<sup>1,3</sup> Scott T. Retterer,<sup>2,4</sup> and **Mitchel J. Doktycz**<sup>1,2,4</sup>

<sup>1</sup>Graduate School of Genome Science and Technology, University of Tennessee-ORNL, Knoxville; <sup>2</sup>Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; <sup>3</sup>Dept. of Biochemistry, Cellular, and Molecular Biology, University of Tennessee, Knoxville; <sup>4</sup>Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, Tenn.

<http://PMI.ornl.gov>

**Project Goals:** See goals for abstract 194.

In order to compete in complex microbial communities, bacteria must quickly sense environmental changes and adjust cellular functions for optimal growth. Such responses are detected and initiated by various two-component signal transduction pathways. Chemotaxis-like signal transduction pathways, which are typically responsible for modulating the swimming motility patterns of chemotactic bacteria, have also been implicated in the modulation of other cellular responses, including cell-to-cell aggregation. Microbial cell-to-cell aggregation is an important behavior which serves to enhance cell survival in adverse environmental conditions. It is particularly advantageous for microbes to organize into aggregative communities not only for protection against predation or antimicrobials, but also for associative metabolic interactions. Plant associative bacteria harbor mechanisms which lead to the remodeling of cell surface and extracellular structures to promote cell-to-cell aggregation and plant microbial interactions. Recently, the Che1 chemotaxis signal transduction pathway from the alpha-proteobacterium *Azospirillum brasilense* was shown to modulate the propensity of cells for cell-to-cell aggregation that leads to flocculation under certain growth conditions. *A. brasilense* are soil diazotrophic bacteria that colonize the roots of many economically important grass and cereal species. Under conditions of high aeration and limiting availability of combined nitrogen, *A. brasilense* cells differentiate into aggregating cells and form dense flocs that are visible to the naked eye. Flocs are formed by cells embedded in a dense polysaccharide matrix and by cell-to-cell aggregation. Optical and electron microscopy (EM) approaches have been previously applied to compare and identify specific cell surface changes that accompany aggregation and flocculation in *A. brasilense* wild-type strain Sp7 and its Che1 pathway mutant strain derivatives that are affected in their propensity to flocculate under nutritional and aeration stresses. However, no specific extracellular structure could be identified using these techniques, despite the preliminary observation

of changes in extracellular polysaccharide (EPS) production detected by growing colonies in the presence of Congo Red. Although optical and EM techniques have revealed many insights into bacterial aggregative behavior, resolution limitations and fixative procedures can inhibit visualization of extracellular structures. Therefore, atomic force microscopy (AFM) was selected as a unique alternative to imaging *A. brasilense* Che1- dependent flocculation behavior at nanometer resolution in an effort to directly visualize changes in cell surface properties that correlate with flocculation. In this study, we investigated *A. brasilense* Sp7 and its Che1 mutant strain derivatives,  $\Delta cheA1$  and  $\Delta cheY1$ , utilizing AFM imaging techniques to gain insight into molecular and regulatory role of Che1 in cell-to-cell aggregation and flocculation. We demonstrate that AFM identifies a distinctive remodeling of the cell surface and extracellular matrix, likely via changes in EPS production, in the  $\Delta cheA1$  and  $\Delta cheY1$  strains concomitant with flocculation under nitrogen-limiting conditions and high aeration.

196

### Plant-Microbe Interfaces: Initial Proteome Characterization of the *Populus* Rhizosphere Community

Chongle Pan<sup>1,2\*</sup> (panc@ornl.gov), Gregory B. Hurst,<sup>2</sup> Robert L. Hettich,<sup>2</sup> Patricia K. Lankford,<sup>3</sup> Manesh B. Shah,<sup>3</sup> Edward C. Uberbacher,<sup>3</sup> Timothy J. Tschaplinski,<sup>4</sup> Sara Jawdy,<sup>4</sup> Gerald A. Tuskan,<sup>4</sup> Lee E. Gunter,<sup>4</sup> Udaya C. Kalluri,<sup>4</sup> Christopher W. Schadt,<sup>3</sup> Neil R. Gottel,<sup>3</sup> Dale A. Pelletier,<sup>3</sup> and **Mitchel J. Doktycz**<sup>3</sup>

<sup>1</sup>Computer Science and Mathematics Division, <sup>2</sup>Chemical Sciences Division, <sup>3</sup>Biosciences Division, and <sup>4</sup>Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

<http://PMI.ornl.gov>

**Project Goals:** See goals for abstract 194.

A complex consortium of microorganisms lives in the plant rhizosphere – a narrow region of soil surrounding the roots. The rhizospheric microbial community feeds on proteins and sugars released by the roots and plant material sloughed off from the roots. In return, the community helps plants to acquire water, nitrogen, and other minerals, and to suppress disease. Here, we describe the development of proteomics methods for understanding this symbiotic relationship. Samples were acquired from both greenhouse-grown and wild *Populus*, the latter including specimens growing in clay and in sandy soil in the Caney Fork State Park in Middle Tennessee (see poster by Schadt et al. for more details). Proteins were extracted from fine roots and associated microorganisms, followed by shotgun proteomics analysis using 2-dimensional liquid chromatography and tandem mass spectrometry. Peptides and proteins were identified by searching mass spectral data against a composite protein sequence database comprised of *Populus* proteins and

genomes of relevant sequenced microorganisms. Because the membership and physiology of the rhizosphere community are influenced by growth conditions and genotypes of *Populus* and environment variables associated with soil, results from these different sites and specimens provided a broad survey of the proteomes of *Populus* rhizosphere community. We identified a core rhizosphere proteome present across all samples and sets of proteins unique to soil types or *Populus* growth conditions. Experiments to improve the depth of the community proteome coverage through optimization of protein extraction methods and employment of high-performance mass spectrometers are ongoing. The observations from these field samples will be followed up with measurements on controlled, reconstituted small soil-less or soil-based model systems. Informatics tools to search proteomic data using the metagenomic sequence data collected from the same samples are also being developed.

## 197

### Plant-Microbe Interfaces: Development of a Knowledgebase for Exploring Plant-Microbe Interactions Using Metabolic Reconstructions

Tatiana Karpinets<sup>1\*</sup> (karpinetsv@ornl.gov), Michael Leuze,<sup>2</sup> Guruprasad Kora,<sup>2</sup> Dale A. Pelletier,<sup>1</sup> Mustafa Syed,<sup>1</sup> Byung Park,<sup>2</sup> Denise Schmoeyer,<sup>2</sup> Ed Uberbacher,<sup>1</sup> and **Mitchel J. Doktycz**<sup>1</sup>

<sup>1</sup>Biosciences Division and <sup>2</sup>Computer Science and Mathematics, Oak Ridge National Laboratory, Oak Ridge, Tenn.

<http://PML.ornl.gov>

**Project Goals: See goals for abstract 194.**

The Plant-Microbe Interfaces knowledgebase provides an integration framework for microbial, fungal and plant genome data along with information about the characteristics of plants, soils and weather conditions from field studies, their geographical location, diversity and the phenotypic characteristics of microorganisms from soil and plant samples at each location. This macro level information on the samples will be later supplemented by proteomics, metabolomics and transcriptomics data. In this study we have utilized the database to explore metabolic signatures for different classes of organisms interacting with *Populus*. In natural environments a tripartite interaction among plants, fungi, and bacteria is essential to plant growth. On one hand, the intimate relationships within the plant/bacteria/mycorrhizal fungus network supplies plants with nutrients, promotes their growth, and increases their resistance to stress. On the other hand, pathogenic fungi and bacteria can be harmful to plants and can lead to diseases and to altered production of desired traits. Additionally, some microorganisms, also known as biocontrol agents, can protect plants by reducing the number of pathogenic microorganisms. Molecular mechanisms underlying phenotypic differences among plant-associated microorganisms are not clear. The

microbial species responsible for beneficial and pathogenic effects on plant growth that have been sequenced provide an opportunity to discover the genomic determinants of the various phenotypes through comparative analysis of the genomes. Metabolic capabilities inferred from genome annotations can be especially helpful in fingerprinting phenotypic differences at the level of enzymes and metabolic pathways. In this study we perform a comparative analysis of phenotypes by developing Pathway Genome Databases (PGDBs) for two fungal species, *Eremothecium gossypii* ATCC 10895 (a plant pathogen) and *Laccaria bicolor* S238N-H82 (a plant symbiont) and a set of sequenced bacterial species including four endophytes (*Pseudomonas putida* W619, *Stenotrophomonas maltophilia* R551-3, *Enterobacter* sp. 638, *Methylobacterium populi* BJ001), six plant pathogens from the genera *Agrobacterium*, *Pseudomonas*, and *Burkholderia*, and five species from the same genera that are used for biocontrol of phytopathogens.

PGDBs were generated by the Pathologic program from the Pathway Tools software. Because the quality of the metabolic reconstruction by this software depends on the genome annotation, we have developed an automated pipeline to improve the enzyme annotation and to make it consistent across studied organisms. The primary input file to the pipeline contains all of an organism's RefSeq files downloaded from the NCBI website. This file is parsed to build input files for Pathologic and to augment them with the enzyme information from the KEGG orthology annotation of the organism's genome. The PGDB is built with Pathologic running in batch mode, and the PGDB is then refined, to predict transcription units and transporters. Additionally, MySQL tables are created to characterize each protein coding sequence in the genome by a product name, EC numbers, pathways, and protein domains. The domain annotations are generated by searching each sequence against a set of databases (CDD, Pfam, SMART, TIGRFAM, and COG) using RPSBLAST. Thus, the pipeline allows us to quickly incorporate the latest annotation information into the KnowledgeBase, supports metabolic reconstructions, provides a means for improving their quality, and facilitates a comparative analysis of the organisms.

A preliminary analysis of the information generated by the pipeline has revealed some interesting metabolic differences between pathogenic and beneficial microbes at the level of specific enzymes and metabolic pathways. Across analyzed microbial species, biological control agents have a significantly larger number (~10-20%) of metabolic enzymes and pathways when compared with either plant pathogens or endophytes ( $p < 0.05$ ). No statistically significant differences, however, were observed when plant pathogens were compared with plant endophytes, indicating a closer metabolic relationship between these phenotypes. In addition to increased metabolic versatility, all of the analyzed biocontrol agents encode in their genomes two specific enzymes that clearly distinguish their metabolic capabilities from the analyzed pathogenic bacteria and endophytes. One enzyme, 1,6-dihydroxycyclohexa-2,4-diene-1-carboxylate dehydrogenase (EC 1.3.1.25), is involved in benzoate degradation, namely in the conversion of benzoate to catechol. It is pres-

ent in biocontrol agents but is not found in any pathogen. The other enzyme, 1-aminocyclopropane-1-carboxylate (ACC) deaminase (EC 3.5.99.7), is absent from endophytes, but is common for biocontrol bacteria. This enzyme catalyzes the conversion of ACC, a precursor of ethylene synthesis in plants, to  $\alpha$ -ketobutyrate and ammonia. A variety of beneficial effects on plant growth has been linked to this enzyme, including enhanced nodulation and increased resistance to stress.

## 198

### Plant-Microbe Interfaces: Novel Navigation Techniques to Study Plant-Microbe Associations Utilizing Google Maps API

Guruprasad Kora<sup>1\*</sup> (koragh@ornl.gov), Tatiana Karpinets,<sup>2</sup> Denise Schmoyer,<sup>1</sup> Michael Leuze,<sup>1</sup> Byung Park,<sup>1</sup> Mustafa Syed,<sup>2</sup> Dale A. Pelletier,<sup>2</sup> Gerald A. Tuskan,<sup>3</sup> Christopher W. Schadt,<sup>2</sup> Ed Uberbacher,<sup>2</sup> and Mitchel J. Doktycz<sup>2</sup>

<sup>1</sup>Computer Science and Mathematics, <sup>2</sup>Biosciences Division, and <sup>3</sup>Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

<http://PMI.ornl.gov>

**Project Goals:** See goals for abstract 194.

Efficient integrative analysis, navigation and visualization of information are important prerequisites to understanding dynamic functional interactions between plants and microbes in the environment. The information that must be collected and analyzed includes a diverse set of data characterizing plants, microbe and their environments at different levels of resolution. The collected data will include characteristics of plants, soils and weather conditions from field studies, their geographical location, diversity and phenotypic characteristics of microorganisms from soil and plant samples at each location. This macro-level sample information will be supplemented by proteomics, metabolomics and transcriptomics data, and by genome annotations of the sequenced microorganisms. These micro-level measurements will be also supplemented by results of statistical analyses and by computational predictions produced by developed mathematical models. The unique diversity of the collected data and their composite nature necessitates a more flexible way of managing the information in the project knowledgebase than commonly used relational databases and related visualization and navigation tools.

To address the complex nature of the collected information in studies of plant-microbe associations, a novel navigation and visualization technique based on Google Maps Application Programming Interface (API) is being developed. This approach will allow us to combine a comprehensive mashup of the collected data, high speed visualization and a facilitated supplementation of the data by available knowledge from public Internet resources. The mashup

tool provides novel navigation techniques to quickly and efficiently locate and zero-in on interesting plant-microbe associations. The web application will integrate data from different sources and explore experimental datasets that are based upon a common geographical and biological sample space. The technology helps users to generate ideas by identifying relationships and associations both within plants and microbes, and between plants and microbes. The tool provides a sophisticated and intuitive graphical user interface to interactively browse the data using different navigation parameters. It provides for a rich interactive user-experience for users to; 1) filter data by biological, geographical or categorical traits of the collected samples; 2) correlate and compare data based on the user selected parameters; and 3) identify interesting trends and patterns in the collected datasets. Particularly, the tool is used to explore a relationship between soil characteristics, plant genotypic and phenotypic characteristics, and microbial phenotypic, metagenomic and metaproteomic characteristics in field-based studies of *Populus* associated microbial communities. The web application is accessible from <http://pmi.ornl.gov>.

## 199

### Plant-Microbe Interfaces: Functional Analysis of Phytochrome Signaling in *Populus*

Abhijit A. Karve<sup>1\*</sup> (karveaa@ornl.gov), David J. Weston,<sup>1</sup> Sara S. Jawdy,<sup>1</sup> Lee E. Gunter,<sup>1</sup> Stan D. Wullschleger,<sup>1</sup> and Mitchel J. Doktycz<sup>2</sup>

<sup>1</sup>Environmental Sciences Division, and <sup>2</sup>Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

<http://PMI.ornl.gov>

**Project Goals:** See goals for abstract 194.

In addition to providing energy for photosynthesis, light also modulates the spatial and temporal responses of plants to their surrounding environment. Plants have evolved multiple mechanisms to modulate their growth and development in response to the spectral quality of light. One such mechanism involves red (R) and far-red (FR) light perception by phytochromes (PHY), where a reduction in R:FR due to vegetative shading results in 'shade avoidance response' (SAR). The SAR is characterized by rapid elongation of stem internodes and upward reorientation of leaves. The role of PHYs in shade avoidance is widely studied in the model plant *Arabidopsis thaliana*. In *Arabidopsis*, phytochromes are encoded by PHYA, PHYB, PHYC, PHYD and PHYE; of these, PHYB plays an important role in mediating responses to plant-plant competition. Here we report initial characterization of the PHY gene family from *Populus trichocarpa*. Unlike *Arabidopsis*, the *Populus* genome encodes three PHY genes namely, PtPHYA, PtPHYB1 and PtPHYB2. In order to gain insight into the role of PtPHYs in light signaling, the transcripts of the three PtPHYs in different plant tissues were measured by quantitative RT-PCR. Consistent

with the light labile nature, PtPHYA transcript was the least abundant of all three PHYs studied. The expression of PtPHYB1 was highest in female flowers and that of PtPHYB2 was highest in the phloem. In *Arabidopsis*, SAR is characterized by upregulation of key response genes such as *ATHB4*, *HFR1* and *PIF3*. *Populus* homologues of *ATHB4* and *HFR1* as well as PtPHYB1 and PtPHYB2 showed significant upregulation in response to low R:FR ratio. These results suggest that the mechanism of PHY signaling is partly conserved between *Arabidopsis* and *Populus*. The gene regulatory networks involved in SAR are being studied by microarray analysis on *Populus* exposed to lower R:FR light. Finally, phytochromes are protein kinases and are believed to affect the downstream responses by interacting with other proteins. In order to identify the PHY interacting proteins, *Populus*-specific PHYs were cloned as C-terminal and N-terminal green fluorescent protein (GFP) and hemagglutinin (HA) tagged constructs. The protein localization was then studied by expressing the GFP-fusion protein in sweet pea leaf mesophyll protoplasts. The candidate PtPHY interacting proteins will be identified by protein-protein interaction assays in leaf mesophyll protoplasts. This work will lead to a conceptual model of phytochrome-mediated responses to shade avoidance and to a more detailed understanding of light-induced signaling cascades in *Populus*.

## 200 Plant-Microbe Interfaces: Transcript and Protein Evidence for Novel Small Protein Genes in *Populus*

Xiaohan Yang<sup>1\*</sup> (yangx@ornl.gov), Gregory B. Hurst,<sup>2</sup> Abhijit A. Karve,<sup>1</sup> Timothy J. Tschaplinski,<sup>1</sup> Sara Jawdy,<sup>1</sup> Patricia K. Lankford,<sup>2</sup> Manesh B. Shah,<sup>4</sup> Gerald A. Tuskan,<sup>1</sup> Lee E. Gunter,<sup>1</sup> Christa Pennacchio,<sup>4</sup> and **Mitchel J. Doktycz<sup>3</sup>**

<sup>1</sup>Environmental Sciences Division, <sup>2</sup>Chemical Sciences Division, and <sup>3</sup>Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn. and <sup>4</sup>DOE Joint Genome Institute, Walnut Creek, Calif.

<http://PMI.ornl.gov>

**Project Goals: See goals for abstract 194.**

Small proteins less than 200 amino acids in length encoded in short open reading frames (sORF) have major functions. Most of the small proteins characterized so far play important roles in cell-to-cell signal transduction. Hundreds or even thousands of novel sORF genes may await discovery in each plant species. The goal of this research is to systematically study the functional genomics of small proteins in relation to signal transduction involved in plant-microbe interactions. We hypothesize that small mobile proteins mediate long-distance signal transduction via the phloem/xylem channels. A systems biology approach is being used to realize three objectives: 1) discovery of sORF candidate genes using bioinformatics, transcriptomics and proteomics;

2) molecular characterization of the candidate genes using full-length gene cloning, sub-cellular localization and movement, and protein-protein interactions; and 3) functional characterization of the candidate genes using signal transduction assays and regulation of gene expression.

Our transcriptomics and proteomics research revealed thousands of sORF genes expressed in *Populus* leaf tissue under normal and drought conditions. Our recent comparative genomics analysis identified some interesting small protein candidates that are potentially involved in signal transduction via the phloem/xylem channels. Proteomics data were also analyzed for the presence of these small proteins. Protein extracts of leaves were prepared, and shotgun proteomics measurements were performed using an LC-MS-MS approach. By comparing experimental tandem mass spectra against a combined database containing current *Populus* gene annotations plus novel small protein candidates, we obtained proof-of-principle confirmation that small proteins are amenable to detection using these protocols. To identify more sORF candidate genes, sequencing of transcriptome and proteome in *Populus* phloem/xylem sap, as well as root system, including rhizosphere microorganisms (e.g., inoculated by fungus *Laccaria*), will be carried out.

For molecular characterization, small protein genes are being cloned as C- or N-terminal GFP (green fluorescent protein) and HA (hemagglutinin) tag constructs. The recombinant protein will be used for in vivo localization assays. The sub-cellular localization of the proteins will be studied by expressing the GFP-tagged constructs in the leaf mesophyll protoplasts. The inter-cellular long-distance movement of small proteins in plants will be monitored by imaging of fluorescent and radiolabeled protein probes. To characterize the role of the small proteins in mediating the plant microbe interactions, candidate genes will be analyzed by making transgenic *Populus* plants and/or hairy roots with altered gene expression (i.e., over-/down-regulation). In addition, the role of the small protein candidates will be studied using protoplast-based signaling assays.

## 201 Plant-Microbe Interfaces: Deciphering Plant-Microbe Signaling with Integrated Networks

David J. Weston<sup>1\*</sup> (westondj@ornl.gov), Andrey Gorin,<sup>2</sup> Yunfeng Yang,<sup>3</sup> Sara S. Jawdy,<sup>1</sup> Abhijit A. Karve,<sup>1</sup> Jennifer Morrell-Falvey,<sup>3</sup> Gerald A. Tuskan,<sup>1</sup> and **Mitchel J. Doktycz<sup>3</sup>**

<sup>1</sup>Environmental Sciences Division, <sup>2</sup>Computer Science and Mathematics, and <sup>3</sup>Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

<http://PMI.ornl.gov>

**Project Goals: See goals for abstract 194.**

The mechanisms by which symbiotic fungi engage host plant systems and induce local and systemic resistance

to pathogens, and promote growth and other beneficial effects remains poorly understood. A goal of our project is to develop experimental and computational methods to discover the signaling cascades and subsequent biochemical and molecular mechanisms driving these mutually beneficial interactions between host plants and fungal symbionts. A systems-level approach will be used to address this goal by: i) constructing a coexpression network map to characterize host plant transcriptional profiles in response to environmental perturbations with and without fungal symbionts; ii) developing computational methods to map regulatory and metabolic pathways implicated in plant-microbe mutualism; and iii) integrating proteome, metabolome, and transcriptome profiles to refine understanding of crucial genes and molecular mechanisms involved in these interactions.

To address these tasks, we have developed a model system where *Arabidopsis thaliana* was grown alone or co-cultivated with the root colonizing mycorrhiza-like fungi *Piriformospora indica*. An abiotic stress coexpression network map was created for *A. thaliana* that consisted of subnetworks (modules) enriched with gene products contributing to signaling, drought, heat, salt, and UV-B perturbations. We hypothesized that these modules would demonstrate robustness to environmental perturbation when *A. thaliana* was co-cultivated with *P. indica*, since this fungus has been shown previously to enhance host plant yield under abiotic stress conditions. After confirming that *P. indica* hyphae were associated with host plant roots using microscopy, plants with and without *P. indica* were subjected to ambient (25 °C) and heat (38 °C) treatments. Regardless of treatment, co-cultivation with *P. indica* induced gene expression in the heat shock and ROS (reactive oxygen species) signaling subnetworks. Co-cultivated plants exposed to the heat treatment maintained other subnetworks similar to ambient temperature expression levels, whereas heat treated plants cultured in the absence of the fungal symbiont had significant expression level changes for all subnetworks relative to ambient temperature controls. Biochemical quantification of ROS confirmed the network-derived signaling result and suggests that *P. indica* manipulates this signaling pathway to buffer host plants against abiotic perturbations.

To further refine our ability to discover biochemical and molecular mechanisms of plant - fungal mutualisms, expression changes in over 200 currently known co-expression modules reflecting tightly coordinated pathways in *A. thaliana* between inoculated and control samples are being systematically investigated. Rearrangements in the co-expression modules in response to mutualistic interactions, specifically genes that are co-expressed in the subnetworks, are being mapped. These genes will be targeted for further analytical studies as potential candidates involved in inter-specific signaling pathways. Finally, a Bayesian statistical framework, calibrated on several well-studied *A. thaliana* molecular mechanisms to integrate proteome, metabolome and transcriptome data, is being used to decipher novel molecular mechanisms corresponding to the co-expression subnetworks that are found to be important in response to interactions with *P. indica*. The overall experimental and

computational approach will be extended to future plant/microbiome communities investigated in this project.

## 202

### Plant-Microbe Interfaces: Characterization of Native Microbial Communities in the Roots and Rhizosphere of *Populus deltoides*

Christopher W. Schadt<sup>1\*</sup> (schadtcw@ornl.gov), Hector F. Castro-Gonzales,<sup>1</sup> Neil Gottel,<sup>1</sup> Dale A. Pelletier,<sup>1</sup> Jennifer Morrell-Falvey,<sup>1</sup> David J. Weston,<sup>2</sup> Rytas Vilgalys,<sup>3</sup> Gerald A. Tuskan,<sup>2</sup> and **Mitchel J. Doktycz**<sup>1</sup>

<sup>1</sup>Biosciences Division and <sup>2</sup>Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn. and <sup>3</sup>Dept. of Biology, Duke University, Durham, N.C.

<http://PMI.ornl.gov>

#### Project Goals: See goals for abstract 194.

The beneficial associations between plants and microbes exemplify complex, multi-organism systems that are shaped by the participating organisms and the environmental forces acting upon them. These plant-microbe interactions can benefit plant health and biomass production by affecting nutrient uptake, influencing hormone signaling, effecting water and element cycling in the rhizosphere, or conferring resistance to pathogens. The rhizosphere of *Populus* species represents an ideal system for understanding the natural diversity of these associations, as well as the molecular details that result in function. *Populus*, and more generally the Salicaceae (willow family) to which they belong, are host to bacterial endophytes and to two prominent types of mycorrhizal fungi, arbuscular endomycorrhizae (AM) formed by Glomalean fungi, and ectomycorrhizae (EM) formed by Ascomycetes and Basidiomycetes.

Broad-based efforts to understand the natural diversity of microbial associates of *Populus deltoides*, *P. trichocarpa*, and their natural and artificial hybrids in both native and controlled habitats have been initiated. As a pilot study for this work, a population of *P. deltoides* as it occurs along the Caney Fork River was sampled in the early fall of 2009. Two *P. deltoides* stands were sampled, representing both upland and bottomland ecotypes and soil conditions that commonly occur in this region. This study is also serving as a “testbed” for methods development that will be applied more broadly for similar studies in the coming growing season. These efforts include four related foci: 1) Community assessment based on rRNA genes using pyrosequencing and other methods to describe the variation within and between individual trees in these two environments (described here); 2) Efficient cultivation, typing and physiological characterization of representative associates (described in poster by Pelletier et al.) that can be used as models for further molecular interaction studies; 3) Methods development for localization and quantification of microbial associates within

diverse *Populus* tissue-types; and 4) Methods development for single-cell manipulations that will compliment targeted cultivation, as well as metagenomic and metaproteomic efforts (described in poster by Retterer et al).

As part of this study, we are conducting 454 based pyrosequencing to describe both bacterial and fungal root and rhizosphere associations of the Caney Fork populations. Rhizosphere populations were examined from field samples that were washed in buffered saline. Mycorrhizal and bacterial tissue-associated populations are being examined on the same surface sterilized root samples. Extensive efforts to optimize both surface sterilization methods and subsequent DNA extraction showed that efficient sterilization could be achieved by combining hydrogen peroxide and sodium hypochlorite based washing followed by commercially available DNA extraction methods. Using these methods, approximately 50-100ng DNA per mg root tissue was extracted and was readily amplifiable with PCR based approaches. Existing methods targeting the V1-V2 region of the 16S rRNA genes have shown that root and rhizosphere associated communities are extremely diverse, comprising thousands of OTUs per sample. Similar to past studies we show that these communities are dominated by alpha-proteobacteria. Comparisons of variation within individual tree samples, between trees from similar environments, and between ecotypes are in process pending the completion of final samples from these collections. Our initial efforts in developing efficient methods for fungal community assessment have focused on descriptions of the D1/D2 region. This region can be targeted with conserved primer sets, allowing alignments over the entire diversity of fungi, but is also variable enough to allow for robust assessment of populations to the family level or below. Initial tests with multiple barcoded primer variants show that efficient and unbiased assessment may be achievable with the developed methods.

## 203

### Plant-Microbe Interfaces: The Role of Plant Genotype and Phenotype in Regulating the Symbiotic Microenvironment

Timothy J. Tschaplinski<sup>1\*</sup> (tschaplinstj@ornl.gov), Udaya Kalluri,<sup>1</sup> Lee E. Gunter,<sup>1</sup> Sara Jawdy,<sup>1</sup> Gerald A. Tuskan,<sup>1</sup> Maud Hinchee,<sup>2</sup> Jesse Labbé,<sup>3</sup> Francis Martin,<sup>3</sup> and Mitchel J. Doktycz<sup>4</sup>

<sup>1</sup>Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; <sup>2</sup>ArborGen LLC, Summerville, S.C.; <sup>3</sup>INRA, Nancy, Champenoux, France; and <sup>4</sup>BioSciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

<http://PML.ornl.gov>

**Project Goals: See goals for abstract 194.**

A number of signals broadly classified as phytohormones (auxin, cytokinin and ethylene) and secondary metabolites

(flavonols, alkaloids and polyamines) have been known to affect plant-microbe associations, however, the molecular underpinnings of how the signals are transduced into plant level phenotypic changes (e.g. lateral root proliferation, induction of systemically-acquired resistance) is poorly understood. Genotypic variability in plant carbon and nitrogen metabolism greatly alters both the intracellular and extracellular metabolite profiles and thus determines the biochemical microenvironment in which microbial (fungal and bacterial) symbionts exist. Our previous biochemical analyses of pure species of black cottonwood (*Populus trichocarpa*), eastern cottonwood (*P. deltoides*), their interspecific hybrids, other *Populus* sp. and their transgenic and cisgenic mutants provide a diverse variety of clones with tailored differences in primary and secondary metabolism that can be exploited to probe plant-microbe interactions. We hypothesize that enhanced production of metabolites of primary C and N metabolism in *Populus* roots, including simple sugars, organic acids, and amino acids, promotes colonization by ectomycorrhizal fungi and endophytic bacteria. We also hypothesize that unique secondary metabolites produced by *Populus* roots function as selective agents that both promote and inhibit specific microbial species. Here we present the current status of *Populus* genetic resources that have been characterized or created through transgenesis to provide the basis for future studies on *Populus*-microbe associations.

*Populus trichocarpa* (T) and *P. deltoides* (D) differ in their profiles of secondary metabolites and these metabolic differences segregate in successive generations such that quantitative trait loci associated with metabolite production (mQTL) can be mapped. We have created dense genetic marker maps and mapped several hundred mQTL for both leaf and root metabolites in two pseudo-backcross (TDxD') pedigrees, including family 13 and family 52124. The identification of 38 definitive and 117 suggestive mQTL for root metabolites in family 13 included the location of an mQTL hotspot on linkage group X that regulates the production of several key secondary metabolites, including salicortin, salireposide, and several phenylpropane glycosides (vimalin-like). Five candidate genes were selected within the mQTL region and constructs were designed for their up- and down-regulated expression. These cisgenic transformants with putatively increased and decreased secondary metabolism are scheduled to be available for characterization early in 2010 and will be available to test the effects of altered metabolism on plant-microbe associations. Additionally, the F2 backcross progeny with extremely high (and low) production of secondary metabolites that led to the identification of 239 mQTL associated with 105 metabolites in family 52124 are available and can also be selected to determine the effects of altered metabolism on plant-microbe associations.

We have shown that a key difference in secondary metabolites among *Populus* species is the nature and concentration of hydroxycinnamate-quininate/shikimate esters that are present and are likely to affect microbial colonization. The role of such metabolites in colonization can be assessed by the selection of *Populus* species with diverse profiles of such metabolites. Additionally, we are generating *P. deltoides*

clones that have been down-regulated for all genes in the lignin biosynthetic pathway, including hydroxycinnamate-quininate transferase (HCT) and coumarate-3-hydroxylase (C3H), that all have direct and indirect effects on the concentrations of these metabolites, given that they serve as storage compounds for the up-stream lignin precursors that have inhibitory effects on microbes. In addition to these clones, we have determined the metabolic phenotype of a number of activation-tagged *P. tremula x alba* clones that have elevated or greatly depleted concentrations of these metabolites. Microbial species that are successful colonizers must be able to tolerate the free mono- and diphenolic acids and their quinate/shikimate esters. These *Populus* clones can serve as the background plant material to determine what classes of compounds promote or inhibit key plant-microbe associations.

We have selected the early colonization events of the *Populus-Laccaria* association, which is the model perennial tree-fungal association with both organisms having had their genomes sequenced and thus have available a broad array of genomic resources. *Laccaria bicolor* is an ectomycorrhizal fungus that routinely colonizes *Populus*. The metabolic signaling responses involved in the establishment of the association is being characterized by analyzing the time-course of metabolomic and transcriptomic (array-based) responses of both organisms reared under *in vitro* culture, contrasted with greenhouse pot culture with samples in both studies collected every 2 weeks up to 8 weeks in the *in vitro* study, and up to 12 weeks in the pot culture study. Additionally, the role that the poplar genotype has in the metabolic responses involved in the establishment of the plant-microbe association is being investigated by altering the *Populus* host, including *P. deltoides*, *P. trichocarpa*, and three *P. trichocarpa x deltoides* hybrids. Once putative signals are identified, their roles will be assessed by initiating assays of indirect contact between *Populus* roots and *Laccaria in vitro* (i.e., a cellophane membrane allows molecular cross-talk through diffusible metabolites, such as auxins, without physical contact) at early time points. Similarly, the effects of the direct contact of a limited number of putative signals at very early time points will also be studied. The metabolite signals and microarray responses will be correlated in co-expression networks to identify novel signaling pathways that regulate the *Populus-Laccaria* association.

Recent reports suggest that root growth induced in *Populus* interacting with *Laccaria* required polar auxin transport as well as auxin signaling through *Populus* auxin response regulator proteins. Many sequenced microbial genomes carry genes of auxin biosynthetic pathway, but it is unclear to what extent these signaling pathways are universal or specific in establishing a symbiotic relationship between a specific microbe and host genotype. We hypothesize that certain auxin response factor proteins belonging to the *Aux/IAA* and *ARF* families play a direct role in establishment and/or signal transduction post-establishment. We are testing this hypothesis by co-culturing specific microbial strains with PCR-confirmed *Populus* RNAi lines specific to genes from the *Aux/IAA* and *ARF* gene families. A micropropagation protocol established to generate whole plantlets from

*Populus* shoot tips is being used to also test candidate endophytic and rhizosphere microbes identified from *Populus* field surveys and will be tracked by imaging and molecular profiling methods. The efforts to identify the molecular factors are being complemented with measurements of the levels and impact on altered levels of auxin in co-cultivation experiments. These plant materials will be harvested at various timepoints and characterized at the transcriptomic and proteomic levels in order to understand the mechanisms of hormone cross-talk relaying the plant level outcomes due to the microbial association. This presents a pipeline for micro-scale screening of where, when, and how microbes associate with the host plant.

## 204 Plant-Microbe Interfaces: Isolation and Characterization of Cultivable Members of the *Populus* Rhizosphere-Endosphere Community

Dale A. Pelletier<sup>1\*</sup> (pelletierda@ornl.gov), Tse-Yuan Lu,<sup>1</sup> Se Yeon Kim,<sup>1</sup> Christopher W. Schadt,<sup>1</sup> Marilyn Kerley,<sup>1</sup> Timothy J. Tschaplinski,<sup>2</sup> David J. Weston,<sup>2</sup> Jennifer Morrell-Falvey,<sup>1</sup> Amy L. Schaefer,<sup>3</sup> E. Peter Greenberg,<sup>3</sup> Caroline S. Harwood,<sup>3</sup> Scott T. Retterer,<sup>1</sup> and Mitchel J. Doktycz<sup>1</sup>

<sup>1</sup>Biosciences Division and <sup>2</sup>Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; <sup>3</sup>Dept. of Microbiology, University of Washington, Seattle

<http://PMI.ornl.gov>

**Project Goals: See goals for abstract 194.**

As part of the newly initiated ORNL Plant-Microbe Interfaces Science Focus area, we are using broad-based methods to characterize the natural diversity of microbial associates of *Populus* and elucidate the molecular mechanisms by which these diverse organisms interact. The goal of the isolation and screening efforts for this project is to develop methods for efficient cultivation, typing, and physiological characterization of representative associates that can be used as model organisms for further molecular interaction studies. In a pilot study for this work, we sampled a population of *P. deltoides* as it occurs along the Caney Fork River in Tennessee in the early fall of 2009. Two *P. deltoides* stands were sampled, representing both upland and bottomland ecotypes and soil conditions that commonly occur in this region. From these samples, a number of diverse fungal and bacterial associates have been isolated from *Populus* rhizosphere and surface sterilized root tissues using broadly compatible media and direct plating methods. These isolates are being screened for phylogenetic identity with bacterial and fungal rDNA primers, and for traits of interest involved in plant-microbe interactions using molecular and biochemical assays, including nitrogen fixation (*nifH*), indole-3-acetic acid synthesis pathways (*iaa*), salicylate degradation genes (*nahJ* and *nahW*), and fungal oxalate production (*oahA*).

The exchange of chemical signaling molecules through either direct contact or diffusion has been demonstrated in a few well studied plant-microbe model systems to effect establishment and maintenance of symbiotic interactions. These signals (e.g., phytohormones, antimicrobials, quorum sensing compounds) affect a wide range of phenotypic responses that can influence plant-microbe and microbe-microbe interactions, including production of antimicrobials, exopolysaccharides, exoenzymes, motility, and conjugation. Microbes isolated from the *Populus* rhizosphere and endosphere are being screened for the production of small signaling molecules by GC-MS metabolomics of culture supernatants from isolates grown in the presence and absence of plant-derived exudates. Isolates are also being screened for known and novel homoserine lactone-derived quorum sensing compounds.

These cultivation and screening efforts are being complemented with improved methods for resynthesizing microbial relationships with *Populus* in greenhouse and tissue culture based assays that are critical components of proving the function of these microbial associates. In addition to conventional plate-based isolation methods, efforts are focused on single-cell microbial isolation and characterization techniques and the development of relevant technologies that will ultimately facilitate the genetic characterization of a greater portion of the endophytic and rhizospheric communities. We anticipate that multiple new microbial species will be characterized and advanced toward further studies.

## 205

### Plant-Microbe Interfaces: Application of Microfluidic Technologies for Microbial Isolation, Cultivation, Characterization, and Emulation of the Plant-Rhizosphere Microenvironment

Scott T. Retterer<sup>1,2\*</sup> (rettererst@ornl.gov), Meena Kalyanaraman,<sup>1</sup> A. Nicole Edwards,<sup>3</sup> Jennifer L. Morrell-Falvey,<sup>1,3</sup> Tim McKnight,<sup>4</sup> Steve L. Allman,<sup>1</sup> James G. Elkins,<sup>1</sup> Gladys Alexandre,<sup>3,5</sup> Martin Keller,<sup>1</sup> and Mitchel J. Doktycz<sup>1,2,3</sup>

<sup>1</sup>Biosciences Division, <sup>2</sup>Center for Nanophase Materials Sciences, and <sup>4</sup>Measurement Science and Systems Engineering Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; <sup>3</sup>Graduate School of Genome Science and Technology, University of Tennessee-ORNL, Knoxville; <sup>5</sup>Dept. of Biochemistry, Cellular, and Molecular Biology, University of Tennessee, Knoxville  
 .....  
<http://PMI.ornl.gov>

#### Project Goals: See goals for abstract 194.

Isolation, cultivation and characterization of microbes from complex samples, derived from soil, rhizosphere and plant material, are essential for defining the plant microbiome.

Technologies that allow the physical and chemical manipulation of fluid volumes that approach the single cell level are critical to such processes and offer a means to understand the full range of genetic and phenotypic variation within plant associated microbial and mycorrhizal communities. Microfluidic platforms, created using soft lithographic and advanced nanofabrication techniques, have been developed and their utility for physically isolating single cells in microbead emulsions and creating complex chemical gradients for the study of microbial taxis has been demonstrated.

Micro-scale alginate beads were created using a microfluidic multiphase droplet generation system. By combining a solution of model cells within the aqueous/alginate phase, the system was successfully used to encapsulate individual and small numbers of GFP-expressing *E.coli*. Both internal and external gelation processes were examined for their effects on cell viability and bead stability. The successful isolation of individual cells within the alginate microbeads and subsequent cultivation of these isolates demonstrates the potential of this system for the small scale clonal expansion of microbial cells captured from more complex mixtures of cells taken from natural samples. As this technology is developed, its implementation in the field and application to the characterization of field samples will play a critical role in assessing the genetic and phenotypic diversity present within the *Populus* microbiome.

Once microbial species are isolated via conventional or emerging cultivation techniques, characterization of their phenotypic response to specific molecular signals found within the plant rhizosphere may be correlated with fluctuations in their population. Such fluctuations may correspond to both developmental and phenotypic stages of plant growth. Understanding microbial taxis in response to components of plant exudate/xylem sap, characterized via mass spectrometry during such stages, may provide insight into what role cell motility and taxis plays on microbial recruitment and colonization. A microfluidic platform that allows exquisite spatial and temporal control of chemical gradients while allowing sustained imaging and tracking of individual bacteria has been developed. This system will be used to examine taxis of model microbes and natural isolates as they become available.

## The Predictive Microbial Biology Consortium

# 206

### Genemap-MS: Stable Isotope Assisted Metabolite Profiling of *Synechococcus sp.* 7002

Richard Baran\* (RBaran@lbl.gov), Benjamin P. Bowen, Steven M. Yannone, and Trent R. Northen

ENIGMA SFA, Lawrence Berkeley National Laboratory, Berkeley, Calif.

<http://www.calccme.org/>

#### Project Goals: Development of a metabolomics-centric platform for comprehensive validation and expansion of genome annotations.

Computational homology-based annotations of sequenced genomes of microbes provide an overview of their metabolic capabilities. Inherent uncertainties in homology-based functional annotations, presence of a significant fraction of genes of unknown function in annotated genomes along with a large number of enzymatic activities without a known corresponding gene limit the extent of genome annotations. Comprehensive profiling of cellular metabolites offers an attractive opportunity for the validation and expansion of genome annotations since the presence of specific metabolites indicates the presence of related enzymatic activities.

The first essential prerequisite for the exploitation of metabolites as indicators of enzymatic activities is the ability to identify these metabolites in complex metabolite profiles. Liquid chromatography coupled to electrospray time-of-flight mass spectrometry (LC-ESI-TOFMS) provides chromatographic separation among metabolites in complex mixtures and mass spectra with high mass accuracy. Metabolites can be identified by comparing their accurate mass, retention time, and fragmentation (MS/MS) spectra against chemical standards. Complex metabolite profiles often contain numerous features which do not correspond to any available and analyzed chemical standard. These features can be analyzed by assigning putative empirical formulas based on accurate mass and isotopic profiles (applicable to small compounds -  $m/z < \sim 200$ ) and inferring partial structural information from MS/MS spectra.

To expand the unambiguous assignment of empirical formulas to larger metabolites (up to  $m/z \sim 500$ ), uniform labeling of *Synechococcus sp.* 7002 cultures was performed with stable isotopes  $^{13}\text{C}$  and  $^{15}\text{N}$ . Characteristic shifts in masses of features in  $^{13}\text{C}$ - and  $^{15}\text{N}$ -labeled metabolite profiles discriminated background signals from metabolites originating from the metabolism of *Synechococcus sp.* 7002. Additionally, these shifts determined the number of carbons and nitrogens in specific metabolites thus discriminating between alternative candidate empirical formulas, which were determined from accurate mass of features in an unlabeled control dataset only. This approach facilitated the

inference of a large number of putative compounds with validated empirical formulas. A fraction of these was assigned to specific metabolites based on the correspondence of their accurate mass and retention time to chemical standard.

Draft metabolic network of *Synechococcus sp.* 7002 was reconstructed using Pathway Tools software (BioCyc). This reconstruction is based on genome annotation but also accounts for additional enzymatic activities or spontaneous reactions inferred from the topology of the metabolic network (gap filling according to reference metabolic pathways). Comparison of the draft network against the set of confirmed empirical formulas showed that many of the unique empirical formulas have no correspondence to any of the 805 metabolites in the draft network. Moreover, a number of the confirmed empirical formulas do not correspond to any metabolite in MetaCyc or KEGG databases.

Following the analysis of MS/MS spectra, the identity of a subset of empirical formulas without any correspondence in MetaCyc or KEGG could be assigned to dipeptides of glutamate (at the N-terminus) with one of multiple hydrophobic amino acids. An intermediate of an alternative biosynthetic pathway, with only one of the pathway's reactions assigned to a gene in the genome, was also identified using MS/MS spectra. Another interesting finding is what appears to be a dead-end metabolite according to KEGG database. This metabolite is present predominantly in the metabolite profile of the culture media, which is significantly less complex than the profile of the cell extract.

Metabolite profiling is an attractive approach for comprehensive interrogation of cellular metabolism. Presence of specific metabolites may serve as an indicator of the presence of specific enzymatic activities or metabolic pathways. Comprehensive metabolite profiling may thus serve as the first step in the validation of genome annotations and identification of candidate enzymatic activities or pathways missing from the genome annotation. Combination of metabolite profiling with genetic/environmental perturbations and transcriptomics/proteomics may further zoom in on specific genes related to specific enzymatic activities or metabolic pathways.

## 207

## Applications of GeoChip for Analysis of Different Microbial Communities

Joy D. Van Nostrand<sup>1,2\*</sup>, Liyou Wu,<sup>1</sup> Patricia Waldron,<sup>1</sup> Ping Zhang,<sup>1</sup> Ye Deng,<sup>1</sup> Zhili He,<sup>1,2</sup> Weimin Wu,<sup>3</sup> Sue Carroll,<sup>4</sup> Chris Schadt,<sup>4,2</sup> Anthony Palumbo,<sup>4</sup> Dave Watson,<sup>4</sup> Craig Criddle,<sup>3</sup> Phil Jardine,<sup>4</sup> Terry C. Hazen,<sup>5,2</sup> and Jizhong Z. Zhou<sup>1,2</sup> (jzhou@ou.edu)

<sup>1</sup>University of Oklahoma, Norman; <sup>2</sup>VIMSS (Virtual Institute of Microbial Stress and Survival) <http://vimss.lbl.gov/>; <sup>3</sup>Stanford University, Stanford, Calif.; <sup>4</sup>Oak Ridge National Laboratory, Oak Ridge, Tenn.; and <sup>5</sup>Lawrence Berkeley National Laboratory, Berkeley, Calif.

**Project Goals:** As part of the GTL program, our research is focused on expanding and improving the GeoChip, a comprehensive functional gene array, and to use this array to detect and monitor microbial communities.

Microarray technology provides the opportunity to identify thousands of microbial genes or populations simultaneously. The objective of this study was to further develop and apply a comprehensive functional gene array (GeoChip) to detect and monitor microbial communities. GeoChip 2.0 is a functional gene microarray which 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. Based on GeoChip 2.0, a new generation, GeoChip 3.0 has been developed, which has several new features. First, GeoChip 3.0 covers more gene groups including antibiotic resistance, energy processing, and additional functional genes involved in C, N, P, and S cycling. Second, the homology of automatically retrieved sequences by key words is verified by HUMMER using seed sequences so that unrelated sequences are removed. Third, a universal standard has been implemented so that data normalization and comparison of different microbial communities can be conducted. Fourth, a genomic standard is used to quantitatively analyze gene abundance. In addition, GeoChip 3.0 includes phylogenetic markers, such as *gyrB*. Finally, a software package has been developed to facilitate management of probe design, data analysis, and future updates. This version of GeoChip was a recipient of a 2009 R&D100 Award from *R&D Magazine*, which recognizes the 100 most innovative scientific and technical breakthroughs of the year. Additional expansion is currently underway and will include genes related to stress response, virulence factors, human-microbiome, and phage genes.

The GeoChip has been used to examine dynamic functional and structural changes in microbial communities from many different environments. Here, examples of studies utilizing the GeoChip to examine microbial communities at contaminated sites are presented. These studies illustrate the ability of the GeoChip to provide direct linkages between microbial genes/populations and ecosystem processes and functions. These 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 using GeoChip 3.0. Functional community dynamics were examined during a period of nitrate exposure. After exposure to nitrate the diversity and richness increased several fold but quickly returned to pre-nitrate levels. Detrended correspondence analysis (DCA) indicated a shift in the overall community structure after nitrate exposure but the community began to return to pre-exposure structure once nitrate was removed. The relative abundance of several nitrogen cycling genes showed an increase immediately after nitrate exposure, including ammonification, denitrification, and nitrogen fixation genes indicating a stimulation of these communities.

(2) In the second study from the FRC, analysis of groundwater monitoring wells along a contamination gradient using GeoChip 2.0 revealed less overlap between wells with different levels of U and NO<sub>3</sub> contamination. While diversity of nitrate-fixation genes decreased in NO<sub>3</sub>-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<sub>3</sub> 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.

(3) A third study is currently underway using GeoChip 3.0 to examine functional gene changes in a U (VI) contaminated area after introduction of a slow-release-substrate (SRS), designed to provide a long-term electron donor for U (VI) reduction. Preliminary results indicate a stimulation of microbial communities. 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.

## 208

**Pipeline for Large-Scale Purification and Identification of *Desulfovibrio vulgaris* Membrane Protein Complexes**

Peter J. Walian<sup>1\*</sup> (PJWalian@lbl.gov), Simon Allen,<sup>2</sup> Lucy Zeng,<sup>1</sup> Evelin Szakal,<sup>2</sup> Eric Johansen,<sup>2</sup> Haichuan Liu,<sup>2</sup> Steven C. Hall,<sup>2</sup> Susan J. Fisher,<sup>1,2</sup> Mary E. Singer,<sup>1</sup> Jil T. Geller,<sup>1</sup> Swan Lin,<sup>1</sup> Terry C. Hazen,<sup>1</sup> H. Ewa Witkowska,<sup>2</sup> **Mark D. Biggin<sup>1</sup>** and Bing K. Jap<sup>1</sup>

<sup>1</sup>Lawrence Berkeley National Laboratory, Berkeley, Calif. and <sup>2</sup>University of California, San Francisco

http://pcap.lbl.gov

**Project Goals: To develop and apply a pipeline for the high-throughput isolation and identification of *Desulfovibrio vulgaris* Hildenborough membrane protein complexes in cultures grown under standard conditions, and to characterize changes in these complexes brought about by environmentally relevant stressors.**

As a component of the LBNL ENIGMA Program, an important task of the Protein Complex Analysis Project (PCAP) is to develop and apply methodologies for the identification of membrane protein complexes isolated from the sulfate reducing bacterium *Desulfovibrio vulgaris* Hildenborough (*D. vulgaris*). Given its demonstrated ability to reduce heavy metals, *D. vulgaris* is projected to play an important role in the management of contaminated sites. To optimize the use of *D. vulgaris* at these sites it will be critical to know how environmental changes affect its performance. To gain insight into these questions, we have developed a pipeline to isolate and identify stable membrane protein complexes present in cultures grown under standard conditions, stationary phase conditions, and in the presence of environmental stressors. Through these efforts we hope to assemble the data needed to characterize stress-induced changes in the relative abundance, composition, and stoichiometry of *D. vulgaris* membrane protein complexes. This data, in turn, will be used to model its stress response pathways and optimize the bioremediation capabilities of this microbe.

Membrane protein complexes pose unique purification and analysis challenges. Largely due to the requirement for detergent solubilization, stable isolation of homogeneous intact membrane protein complexes typically requires separation conditions that are different from those used for water soluble proteins. For this task we have been employing a “tagless” strategy optimized for purifying membrane proteins and then identifying them by mass spectrometry (MS). As opposed to strategies employing affinity tags for the purification of target molecules, use of a tagless strategy will enable us to obtain global views of stress-induced changes involving membrane proteins in *D. vulgaris* cultures grown under a variety of conditions.

In the pipeline, *D. vulgaris* cell membranes isolated from large-scale (100 liter) cultures are first treated with

a relatively mild detergent suited for the extraction of inner-membrane proteins. The residual membranes of this gram-negative bacterium are subsequently treated with a second more active detergent to solubilize proteins of the outer-membrane. Each membrane extract is then independently processed. To purify candidate complexes of the inner- and outer-membrane fractions, ion exchange (IEX) and molecular sieve chromatography are used. Fractions obtained from these procedures are further analyzed using SDS and blue native gel electrophoresis to isolate candidate complexes and obtain molecular weight estimates. To prepare samples suitable for MS analysis, whole lanes are cut from blue-native PAGE gels, placed horizontally along the stacking sections of denaturing gels and subjected to a second dimension of SDS PAGE. Potential complex subunits manifest themselves as bands or spots providing insight into the composition of the native complex. Spots removed from these gels are subjected to in-gel digestion and analysis by liquid chromatography electrospray ionization tandem mass spectrometry (LC-ESI-MS/MS) or liquid chromatography and matrix assisted laser desorption time of flight mass spectrometry (LC-MALDI-MS/MS). Protein identification is achieved by searching a custom *D. vulgaris* database using the Mascot or Protein Pilot search engine.

We are currently completing a baseline membrane protein complex dataset derived from *D. vulgaris* large-scale cultures grown under standard conditions (mid-log phase). In addition to providing a catalog of *D. vulgaris* membrane protein complexes, this data will serve as an essential reference for the detection and characterization of changes in the complexes of cultures subjected to the aforementioned stressors.

## 209

**High Throughput Identification, Purification and Structural Characterization of Soluble Protein Complexes in *Desulfovibrio vulgaris***

Bong-Gyoon Han<sup>1\*</sup> (BGHan@lbl.gov), Haichuan Liu<sup>2\*</sup> (haichuan.liu@ucsf.edu), Ming Dong<sup>1\*</sup> (MDong@lbl.gov), Maxim Shatsky,<sup>1,3</sup> Steven E. Brenner,<sup>1,3</sup> Pablo Arbelaez,<sup>3</sup> Jitendra Malik,<sup>3</sup> Dieter Typke,<sup>1</sup> Terry C. Hazen,<sup>1</sup> Jil T. Geller,<sup>1</sup> Harry J. Sterling,<sup>3</sup> Lee Yang,<sup>1</sup> Megan Choi,<sup>1</sup> Evelin D. Szakal,<sup>2</sup> Simon Allen,<sup>2</sup> Steven C. Hall,<sup>2</sup> Susan J. Fisher,<sup>1,2</sup> Evan R. Williams,<sup>3</sup> John-Marc Chandonia,<sup>1</sup> Jian Jin,<sup>1</sup> H. Ewa Witkowska,<sup>2</sup> Robert M. Glaeser,<sup>1</sup> and **Mark D. Biggin<sup>1</sup>**

<sup>1</sup>Lawrence Berkeley National Laboratory, Berkeley, Calif.; <sup>2</sup>University of California, San Francisco; and <sup>3</sup>University of California, Berkeley

**Project Goals: One of the main goals of the PCAP component of the LBNL ENIGMA SFA is to develop high throughput pipelines to purify and identify protein complexes and structurally characterize them by EM. *Desulfovibrio vulgaris* was selected as a model bacterium because of its bioremediation potential in immobilizing**

toxic heavy metals in soils. Our overall workflow consists of (i) a multidimensional separation of the soluble proteome using tagless strategy; (ii) identification of putative complexes by quantitative mass spectrometry (MS); (iii) further purification of selected complexes; and (iv) structural analysis by EM. In some cases, the stoichiometry of protein complexes is studied using novel mass spectrometry techniques that enable analysis of non-covalent assemblages (native MS). Our major focus is to increase the throughput of MS and EM single-particle analyses and to develop methods for an automated assignment of protein complex components. Currently we are implementing automated collection of EM data, particle boxing, and analysis of structural variation, and the engineering of new support-film technologies for better EM sample preparation to preserve quaternary structure in a conformationally homogeneous state. In the future, we will also take advantage of the SAXS expertise within ENIGMA and incorporate this technology into our structural characterization of complexes.

One of the main goals of the PCAP component of the LBNL ENIGMA SFA is to develop high throughput pipelines to purify and identify protein complexes and structurally characterize them by EM. *Desulfovibrio vulgaris* (*DvH*) was selected as a model bacterium because of its bioremediation potential in immobilizing toxic heavy metals in soils at the DOE sites by rendering them practically insoluble upon conversion to the low red-ox state. Our overall workflow consists of (i) a multidimensional separation of the soluble proteome; (ii) identification of putative complexes by quantitative mass spectrometry (MS) followed by bioinformatics analysis; (iii) further purification of selected complexes; and (iv) structural analysis by EM. In some cases, the stoichiometry of protein complexes is studied using novel mass spectrometry techniques that enable analysis of non-covalent assemblages (native MS). In future, we will also take advantage of the SAXS expertise within ENIGMA and incorporate this technology into our structural characterization of complexes.

To identify protein complexes, we have introduced and established a tagless strategy, which is based on the premise that the great majority of stable protein complexes will survive intact separation through a series of orthogonal chromatographic methods. Under this scenario, co-migrating polypeptide components of protein complexes generate overlapping elution profiles at each stage of protein separation. Changes in the relative concentration of each polypeptide (i.e., their elution profiles across the final chromatographic step) are measured with the aid of isotopic dilution mass spectrometry (MS) and iTRAQ reagents (Dong et al., 2008). Various data analysis approaches are being developed to automate assignment of the identified polypeptides to putative complexes. To date, ~64% of the target protein complex separation space has been analyzed, resulting in the identification of over 900 polypeptides. The great majority of the polypeptides are engaged in intermolecular interactions, as evidenced by more than 70% demonstrating significantly higher elution volume (at least 2x) in size exclusion chromatography (SEC) than expected from the

molecular weight of the polypeptide predicted from genome sequence. In addition to heteromeric complexes, the tagless strategy allows detection of homomers, which are not easy to recognize by other methods (e.g., TAP). Overall, at least 45 heteromeric and over 550 homomeric complexes have been identified so far. To address the challenges posed by a co-elution of non-related polypeptides with the legitimate components of complexes, we are evaluating monitoring elution profiles at the two final stages of protein complex separation (hydrophobic interaction chromatography and SEC), as opposed to the single step (SEC) used so far. The resulting 2D polypeptide elution map is expected to provide higher resolution data and consequently to increase confidence in protein complex assignments.

To determine molecular structures, we selected 16 complexes identified by the tagless strategy with molecular weights 400 - 1,000 kDa and subjected them to single-particle EM analysis. Half of the complexes studied proved stable enough to produce high-quality 3-D reconstructions with a resolution of ~2 nm (Han et al., 2009). This success rate for obtaining structures is about 10 times greater than that of previous "proteomic" screens. We have found that there are a surprisingly large number of differences in the quaternary structures of complexes isolated from *DvH* compared to those of homologous proteins from other microbes. These differences occur so frequently that structures determined for complexes in other micro-organisms are likely to be inadequate as templates for modeling the biochemical networks within a given microbe of interest. By extension, we suspect that it may also be the case that complexes change structure frequently under different physiological conditions and future work will address this possibility. Our major focus now is to increase the throughput of EM single particle structural analysis. This effort currently includes the implementation of automated data collection, particle boxing, and analysis of structural variation (Shatsky et al., 2009), and the engineering of new support-film technologies for better EM sample preparation to preserve quaternary structure in a conformationally homogeneous state.

## References

1. Dong *et al.*, 2008, A "tagless" strategy for identification of stable protein complexes genome-wide by multidimensional orthogonal chromatographic separation and iTRAQ reagent tracking. *J Proteome Res.* 7:1836-49.
2. Han *et al.*, 2009. Survey of large protein complexes in *D. vulgaris* reveals great structural diversity. *PNAS.* 106:16580-16585.
3. Shatsky *et al.*, 2009. A method for the alignment of heterogeneous macromolecules from electron microscopy. *J. Struct. Biol.* 166: 67-68.

## 210

**Protein Complex Analysis Project (PCAP): Large-Scale Identification of Protein-Protein Interactions in *Desulfovibrio vulgaris* Using Tandem-Affinity Purification**

Swapnil Chhabra,<sup>1</sup> Gareth Butland,<sup>1\*</sup> Dwayne Elias,<sup>2</sup> Sonia Reveco,<sup>1</sup> Veronica Fok,<sup>1</sup> Barbara Gold,<sup>1</sup> Thomas Juba,<sup>2</sup> John-Marc Chandonia,<sup>1</sup> Ewa Witkowska,<sup>3</sup> Terry Hazen,<sup>1</sup> Judy Wall,<sup>2</sup> Jay Keasling,<sup>1,4</sup> and **Mark Biggin**<sup>1</sup> (mdbiggin@lbl.gov)

<sup>1</sup>Lawrence Berkeley National Laboratory, Berkeley, Calif.; <sup>2</sup>University of Missouri, Columbia; <sup>3</sup>University of California, San Francisco; and <sup>4</sup>University of California, Berkeley

**Project Goals: Develop a high-throughput platform for chromosomal modifications of microbes of interest to DOE. Identify protein-protein interactions in the model sulfate reducer *Desulfovibrio vulgaris* using tandem affinity purification. Develop computational tools and models to characterize sulfate-reducing bacteria.**

Most cellular processes are mediated by multiple proteins interacting with each other in the form of multi-protein complexes and not by individual proteins acting in isolation. One of the major goals of LBNL ENIGMA SFA is to develop a comprehensive knowledgebase of protein complexes and protein-protein interactions (PPI) in microbes of interest to DOE. As part of the Protein Complex Analysis Project component of ENIGMA, *Desulfovibrio vulgaris*, a sulfate reducing bacterium (SRB) found to exist in several DOE waste sites, has been used as a model to screen for PPI using a high throughput tandem affinity purification (TAP) approach. The challenges in working with *D. vulgaris* are typical of organisms of interest to DOE. While likely the best characterized SRB, *D. vulgaris* is an obligate anaerobe and possessed very limited genetic tools. Our approach therefore required the development of a high throughput pipeline to enable the creation of a library of genetically engineered strains, which builds upon generic principles of recombination of non-replicating gene replacement constructs ("suicide" constructs). We have applied this pipeline to the creation of *D. vulgaris* strains bearing TAP-tagged alleles designed for tandem affinity purification of individually tagged bait proteins; however, the versatility of this approach enables its potential application for chromosomal modifications of the majority of microbes of interest to DOE.

Previously we reported the development of large-scale single- and double- crossover chromosomal integration platforms for generating TAP-tagged strains of *D. vulgaris*. Transformation of plasmids into *D. vulgaris* is inefficient and multiple strategies, including constructing a restriction endonuclease host strain mutant that enhanced transformation of stable plasmids into *D. vulgaris* have been explored. However the single greatest improvement in transformation

and integration of suicide constructs has resulted from the transition from the TOPO-Gateway<sup>®</sup> scheme to the Sequence and Ligation Independent Cloning (SLIC) technique for suicide construct generation. Success rates for suicide construct generation improved from 50% to more than 80%, transformation and integration of constructs into *D. vulgaris* from 34% to 65% and ~60 % of these isolates were found to express a TAP-tagged fusion protein detected by IP-western. Currently, we have a library of over 700 suicide plasmid constructs which have been employed for the generation of over 300 tagged SRB strains of which ~200 baits have been analyzed by TAP to date.

Here we present details of our high throughput pipeline for strain engineering along with results of the subsequent TAP analysis of these engineered strains. The interactions presented cover a range of biological processes, including energy conservation (ATP Synthase, Ech hydrogenase), sulfate reduction (dissimilatory sulfite reductase, adenylylsulfate reductase) and protein secretion (YajC-HflCK complex), and include both novel and previously predicted interactions.

## 211

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

Sung Ho Yoon<sup>1\*</sup> (syoon@systemsbiology.org), Christopher Bare,<sup>1</sup> David Reiss,<sup>1</sup> Dan Tenenbaum,<sup>1</sup> Min Pan,<sup>1</sup> Joseph Slagel,<sup>1</sup> Sujung Lim,<sup>2</sup> June Burn,<sup>2</sup> John A. Leigh,<sup>2</sup> Murray Hackett,<sup>3</sup> Angeli Lal Menon,<sup>4</sup> Michael W.W. Adams,<sup>4</sup> Sunia A. Trauger,<sup>5</sup> Gary Siuzdak,<sup>5</sup> Steven M. Yannone,<sup>6</sup> Benjamin Bowen,<sup>6</sup> Stephen R. Holbrook,<sup>6</sup> John A. Tainer,<sup>6</sup> and **Nitin S. Baliga**<sup>1</sup> (nbaliga@systemsbiology.org)

<sup>1</sup>Institute for Systems Biology, Seattle, Wash.; <sup>2</sup>Dept. of Microbiology and <sup>3</sup>Dept. of Chemical Engineering, University of Washington, Seattle; <sup>4</sup>Dept. of Biochemistry and Molecular Biology, University of Georgia, Athens; <sup>5</sup>Center for Mass Spectrometry, Scripps Research Institute, La Jolla, Calif.; <sup>6</sup>Dept. of Molecular Biology, Lawrence Berkeley National Laboratory, Berkeley, Calif.  
<http://gaggle.systemsbiology.net/projects/doe-archaea/2007-04/>  
<http://maggie.systemsbiology.net>  
<http://baliga.systemsbiology.net>

**Project Goals: Bolster through high-end state-of-art systems approaches, developed specifically for the study of archaeal organisms, the comprehensive analysis of multi-protein complexes in DOE-relevant organisms.**

Rational re-engineering of biology for the purpose of bioremediation, bioenergy or C-sequestration requires deep understanding of all functional interactions of relevant components within native cell (s). Many of these functional interactions are conserved across diverse species to different

degrees depending on their evolutionary distance. We are conducting integrative analysis of genomic architecture and composition, transcriptome and proteome structure/function, protein-protein and protein-DNA interactions and metabolic networks to find keystone complexes and specialized circuit architectures for important application-relevant genes within four archaeal organisms. These organisms have enormous potentials from the standpoint of H<sub>2</sub> production, N<sub>2</sub> fixation, and C-sequestration; they include an anaerobic thermophile (*Pyrococcus*), an acidophilic and aerobic thermophile (*Sulfolobus*); a hydrogenotrophic methanogen (*Methanococcus*), and a photoheterotrophic halophile *Halobacterium* NRC-1. A key aspect of our approach is to use the power of systems biology to delineate the process of nucleation, assembly, and turnover of key complexes. Here, we report comparative analysis of dynamically changing transcriptome structures of the four archaea with special emphasis on the conditional activation of unconventional transcriptional promoters within conserved genes and operons.

**Note:** Computational and experimental results from this study will be freely available upon publication at <http://magie.systemsbiology.net/>. All of the software tools developed in this project have been made freely available at <http://gaggle.systemsbiology.net/projects/doe-archaea/2007-04/>.

This research was supported by U.S. Department of Energy, Award No. DE-FG02-07ER64327 and DG-FG02-08ER64685.

## 212

### Metabolic Transformations and Chemical Differences

Ben Bowen\* (BPBowen@lbl.gov), Richard Baran, Steve Yannone, John Tainer, and **Trent Northen**

ENIGMA SFA, Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.

<http://www.calccme.org>

**Project Goals: Use autocorrelation for the analysis of metabolites using liquid chromatography coupled to mass spectrometry.**

Defining the complete metabolic reconstruction for even a simple organism requires detailed knowledge of enzyme function, activity, and biochemistry. Uncharacterized enzymes and unknown metabolic pathways limit the utility of inferring metabolic capability based on genome sequencing and annotation. Homology-based protein function predictions are limited by evolutionary processes that result in conserved domains, and the complexity of biochemistry is so diverse that only a small portion has been defined. Therefore, improving genome annotations and furthering our knowledge of biochemistry is of tremendous importance to take advantage of the potential of genome sequencing.

The detection and analysis of metabolites using liquid chromatography coupled to mass spectrometry (LC/MS)

has the potential to define metabolic pathways, understand the regulation of substrate utilization, measure biomass composition, and much more. The primary challenge associated with this analytical platform is the interpretation of the thousands of molecular features detected in a typical dataset. There are three hurdles which define this crux: 1) a large fraction of the detected features are associated with uncharacterized metabolites, 2) standard methods for the detection of known metabolites are not widespread, and 3) optimum methods have not been defined for feature identification from raw data.

Although the scale biochemistry is vast, the emergence of primary and secondary metabolites is due to a limited number of elementary chemical-differences relating compounds. In this work, we explore a finite number of chemical transformations, and examine chemical-difference-space to understand alterations in metabolism. This is accomplished using autocorrelation, an established method in data analysis with a very high sensitivity for detecting correlated relationships. Here, we have applied autocorrelation to interpret raw mass spectrometry data and then linked the global correlation spectra to chemical differences. This has enabled a global analysis of LC/MS data without relying on feature identification. This analysis was applied to *Sulfolobus sulfataricum*, an archaea that is of interest in creating biofuels, and early results indicate that discrete chemical differences can be measured and interpreted.

## 213

### Towards Localization of Functionality in *Desulfovibrio vulgaris* by Electron Microscopy

David A. Ball,<sup>1</sup> Swapnil Chhabra,<sup>1</sup> Dwayne Elias,<sup>2</sup> Veronica Fok,<sup>1</sup> Jil T. Geller,<sup>1</sup> Amita Gorur,<sup>1</sup> **Terry C. Hazen**,<sup>1</sup> Danielle Jorgens,<sup>1</sup> Thomas Juba,<sup>2</sup> Ambrose Leung,<sup>1</sup> Jonathan Remis,<sup>1</sup> Mary E. Singer,<sup>1</sup> Andrew Tauscher,<sup>1</sup> **Judy Wall**,<sup>2</sup> **Manfred Auer**,<sup>1</sup> and **Kenneth H. Downing**<sup>1\*</sup> (khdowning@lbl.gov)

<sup>1</sup>Lawrence Berkeley National Laboratory, Berkeley, Calif. and <sup>2</sup>University of Missouri, Columbia

<http://PCAP.LBL.GOV>

**Project Goals: The twin goals of this project are 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.**

Bacteria display a spatially and temporally defined three-dimensional organization of their macromolecular complex inventory that allows them to grow, divide, and respond to

environmental cues. As part of the PCAP component of the LBNL ENIGMA Program, our overall goal is to develop tools that can be used to define the localization of such complexes within the cell.

*Desulfovibrio vulgaris* Hildenborough (DvH) is an anaerobic sulfate-reducing bacterium (SRB). It is used as a model organism for the study of environmental bioremediation of heavy metal and radionuclide contamination. Its ability to efficiently reduce toxic heavy metals such as uranium and chromium is of particular interest to the DOE for use in high-risk metal contaminated sites, and it can also provide novel insights on the range of metabolic pathways available in microbes. Protein expression levels and subcellular localization of proteins and macromolecular complexes may change in response to various environmental factors such as exposure to the target toxins, and may also differ among bacterial cells of identical genetic origin within a given culture. We aim to use our localization technology to understand how protein abundance and spatial distribution relate to the cell's normal metabolism and how changes in these factors are involved in the cell's response to changes in the environment.

For some of the very large complexes that display a distinctive size and shape, structural approaches for identification will allow complex localization, for example by identifying single-particle-EM derived 3D structures in whole mount cryo-tomograms. A high-resolution single particle analysis of the highest molecular weight protein structures purified from DvH cells has been conducted by the PCAP EM project, characterizing structures of 16 individual molecular machines at a resolution compatible with electron tomography (Han et al., PNAS 2009). It was hoped that these 'templates' could be used to localize and monitor the differing numbers of single particle structures within the tomographic datasets of intact DvH cells at distinctive stages of its growth and under different stress conditions.

While our cryo-electron tomography studies of DvH have been quite productive in characterizing aspects of overall morphology, they have so far yielded less information about cytoplasmic composition than we had initially hoped. The resolution of the tomograms within the DvH cytoplasm is noticeably lower than the resolution obtained in tomographic datasets from cells of other microbes of similar size. It was not clear if this lack of information is due to a lower number of internal molecular machines, i.e. ribosomes, GroEL, RNA polymerase etc, within the DvH cells, or if the cytoplasm of DvH is for some unknown reason idiosyncratically more electron dense than in other cells, which would decrease the contrast and thus limit our ability to identify cytoplasmic complexes.

The fledgling technique of vitreous cryo-sectioning has been adopted to elucidate this problem. A number of various cell types, including DvH, yeast, *E. coli* and *Caulobacter* have all been successfully high-pressure frozen in specialized copper tubes, sectioned and imaged in the frozen-hydrated state. Variations in the density of the extracellular medium, such as with dextran that is often used as a cryo preservative, provide a mechanism of contrast matching to complement

other measures of cell mass density. A number of tentative tomographic datasets have been recorded from the sections, and data comparison is ongoing.

For smaller proteins, where shape or size is no longer a unique identifier, we can apply tag-based labeling approaches to identify the precise location of the tagged protein. To this end we have explored the SNAP-tag labeling system for DvH. This approach overcomes problems of using GFP derived-fusion proteins in an obligate anaerobe. Labeled cells can be examined at the light microscope level, and can also be photoconverted to provide contrast for EM visualization at much higher resolution. Over the last year we have overcome several obstacles that had plagued us in earlier work, and we recently were able to fluorescently label 13 out of 25 strains. Successful labeling was judged by fluorescence microscopy, SDS gel electrophoresis and plate reader analysis. Using photoconversion of fluorescent signals, followed by resin-embedding, we have begun to map out a variety of proteins at higher resolution. Several of the fluorescent cell lines have been successfully photoconverted and imaged at the TEM level to identify high-precision subcellular location of the tagged proteins. Interestingly we found cell-to-cell differences in labeling signal strength, which we attribute to real differences in protein expression. We also found cell-to-cell differences in extracellular metal reduction activity.

Energy Dispersive X-ray Analysis (XEDS) has also been used to track the evolution of internal and external electron dense material, which becomes visible through the life cycle of an anaerobic DvH culture. Large internal elemental sulfur balls as well as both internal and external Iron sulfide bodies have been identified and characterized.

Ultimately, in addition to describing the distribution of protein complexes within the cell, by comparing cell-to-cell differences in abundance and subcellular localization with uneven extracellular metal reduction activity we hope to be able to gain insight into the distribution and hence function of candidate proteins, assumed to play a role in extracellular metal reduction and other aspects of the cell's redox chemistry.

## 214

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

**Terry C. Hazen**<sup>1,2\*</sup>, Gary Anderson,<sup>1,2</sup> Sharon Borglin,<sup>1,2</sup> Eoin Brodie,<sup>1,2</sup> Steve van Dien,<sup>8</sup> Matthew Fields,<sup>1,7</sup> Julian Fortney,<sup>1,2</sup> Jil Geller,<sup>1,2</sup> E. Hendrickson,<sup>5</sup> Kristina L. Hillesland,<sup>1,6</sup> Hoi-Ying Holman,<sup>1,2</sup> J. Leigh,<sup>1,6</sup> T. Lie,<sup>1,6</sup> Dominique Joyner,<sup>1,2</sup> Romy Chakraborty,<sup>1,2</sup> Dwayne Elias,<sup>1,3</sup> Aindrila Mukhopadhyay,<sup>1,2</sup> Christopher Schadt,<sup>1,3</sup> David Stahl,<sup>1,6</sup> Sergey Stolyar,<sup>1,6</sup> Chris Walker,<sup>1,6</sup> Judy Wall,<sup>1,5</sup> Zamin Yang,<sup>1,3</sup> Huei-che Yen,<sup>1,5</sup> Grant Zane,<sup>1,5</sup> and Jizhong Zhou<sup>1,9</sup>

<sup>1</sup>Virtual Institute of Microbial Stress and Survival; <sup>2</sup>Lawrence Berkeley National Laboratory, Berkeley, Calif.; <sup>3</sup>Oak Ridge National Laboratory, Oak Ridge, Tenn.; <sup>4</sup>Diversa, Inc.; <sup>5</sup>University of Missouri, Columbia; <sup>6</sup>University of Washington, Seattle; <sup>7</sup>Montana State University, Bozeman; <sup>8</sup>Genomatica, San Diego, Calif.; and <sup>9</sup>University of Oklahoma, Norman

**Project Goals: The environmental microbiology core of the ESPP project is the source of environmental data and samples that determine the stressors that will be studied, provides the environments for growing the organisms to be tested, simulates stressed environments, and verifies the conceptual models to determine how these stress regulatory pathways control the biogeochemistry of contaminated sites.**

### Field Studies

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 Hanford 100H as part of our ERSP project we injected 40 lbs of HRC (polylactate) as a slow release electron donor in August 2004. Until March 2008 reducing conditions were maintained, along with undetectable levels of Cr (VI) (Hubbard et al, 2008; Faybishenko et al., 2008). During this time the environment was dominated by sulfate reducers and we were able to detect *Desulfovibrio vulgaris*-like organisms with direct fluorescent antibody. Last year we injected 10 lbs of HRC at the same site to determine if there is a 'memory' response and observed H<sub>2</sub>S production after only 23 h. Once the community stabilizes we will begin push-pull stress tests in the field with NO<sub>3</sub> and monitor functional gene, community structure, and stress responses as compared with previously published models by our group with pure cultures. We are also isolating consortia and determining the dominant community structure to compare with our lab studies.

In order to characterize microbial community dynamics associated with Cr (VI) biostimulation at the Hanford 100-H area, both groundwater and sediment communities are being tracked via SSU rRNA gene sequences. Stainless steel-mesh columns packed with Hanford sediments are placed down-well and microbial communities are allowed to establish. Four wells are being monitored, and baseline communities have been analyzed for the sediment-associated populations. Predominant populations include: *Acetivibrio*; *Geobacter*; *Ruminococcus*, *Alkaliflexus*, *Thiohalomonas*, *Acidovorax*, *Aquaspirillum*, and *Herbaspirillum*. Groundwater sample analysis is in progress, and biomass loads appear to be lower for groundwater compared to sediments. Rarfaction curves indicate that coverage has not been saturated; therefore, pyrosequencing efforts are also underway. Once baseline groundwater and sediment populations are determined during stimulation with HRC, community dynamics will be tracked during an in situ nitrate perturbation. Methanogen enrichments have yielded slow-growing cultures. A hydrogenotrophic enrichment is dominated by *Methanocella paludicola* (90% similar), and an acetoclastic enrichment is predominated by *Methanosarcina barkeri* (99% similar).

We are also using the GeoChip 3.0, a comprehensive functional gene array contains ~25,000 probes covering ~47,000 sequences for 292 gene families involved in the geochemical cycling of C, N, and S, metal reduction and resistance, and organic contaminant degradation was used to examine microbial communities both at Hanford and at Oak Ridge within a pilot-scale test system established for the biostimulation of U (VI) reduction in the subsurface by injection of ethanol. Sediment from eleven different sampling wells, representing two different treatment zones within this system, was evaluated. The results showed that different microbial communities were established in different wells and high gene overlap was observed from wells within the same treatment zone. Higher microbial functional gene number, diversity and abundance were observed within the active bioremediation zone. The microbial community structure was highly correlated with the hydraulic flow rate and geochemical conditions of the treatment zone, especially pH, manganese concentration and electron donor level. In a different study of the same system, functional community dynamics were examined during a period of oxidation by nitrate. After exposure to nitrate the diversity and richness increased several fold but quickly returned to pre-nitrate levels. The relative abundance of several nitrogen cycling genes showed an increase immediately after nitrate exposure, including ammonification, denitrification, and nitrogen fixation genes indicating a stimulation of these communities after nitrate exposure.

*Desulfovibrio* spp. and consortia genomes. *Desulfovibrio* FW1012B was isolated from contaminated groundwater during biostimulation for U (VI) reduction in Oak Ridge, TN. Genome sequencing of isolate *Desulfovibrio* FW1012B, tentatively *Desulfovibrio oakridgensis*, was completed in late spring of 2009 at the DOE's Joint Genome Institute at Lawrence Berkeley National Lab in conjunction with Los Alamos National Lab, Oak Ridge National Lab, and Montana State University. The G+C content of the chro-

mosome is 66.5%. Based on the draft, incomplete sequence, automated gene scanning software predicted a total of 3,737 protein-encoding genes, 3,190 of which match best to KEGG-annotated genes in Proteobacteria genomes. The closest organisms with fully sequenced genomes are *Desulfovibrio vulgaris* Miyazaki (dvm) at 740 hits, *Desulfovibrio desulfuricans* G20 (dde) at 479 hits, *Desulfovibrio vulgaris* DP4 (dvl) at 375 hits and *Syntrophobacter fumaroxidans* (sfu) at 196 hits. A further 333 predicted protein-encoding genes had no match which could signify novel and/or unique genes. Analysis of the data raises some interesting questions regarding the metabolism of the organism. Unlike most *Desulfovibrio* species, *D. oakridgensis* appears to have a fully intact glycolysis pathway, a pyruvate dehydrogenase, and a nearly intact citric acid cycle.  $\alpha$ -ketoglutarate is made by an NADP<sup>+</sup> dependent isocitrate dehydrogenase rather than NAD<sup>+</sup> dependent which creates and lacks a putative malate dehydrogenase. The presence of nitrogenase supports our observations that the organism can assimilate atmospheric nitrogen, and the lack of an aldehyde dehydrogenase coincides with the inability to use ethanol as electron donor and carbon source. Still, other questions are left unanswered. We are also sequencing a number of Hanford isolates that show association with *Desulfovibrio* at Hanford: *Geobacter metallireducens*, *Pseudomonas stutzeri*, and *Desulfovibrio vulgaris*.

This work was part of the ENIGMA SFA and the Virtual Institute for Microbial Stress and Survival (<http://VIMSS.lbl.gov>) supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy.

## 215

### Laboratory Models for the Study of Community Interaction, Functional Stability, and Survival

**A. Arkin**<sup>1\*</sup> (aparkin@lbl.gov), E. Baidoo,<sup>1</sup> P. Dehal,<sup>1</sup> D. Elias,<sup>2</sup> M. Fields,<sup>3</sup> J. Geller,<sup>1</sup> T. Hazen,<sup>1</sup> Z. He,<sup>8</sup> K. Hillesland,<sup>4</sup> J. Keasling,<sup>1</sup> C. Keller, M. Keller,<sup>5</sup> L. Krumholz, B. Meyer, L. Miller,<sup>6</sup> J. Mosher,<sup>5</sup> A. Mukhopadhyay,<sup>1</sup> A. Palumbo,<sup>5</sup> T. Phelps,<sup>5</sup> M. Podar,<sup>5</sup> L. Rajeev, A. Redding,<sup>1</sup> C. Schadt,<sup>5</sup> D. Stahl,<sup>4</sup> S. Stolyar,<sup>7</sup> A. Venkateswaren, C. Walker,<sup>4</sup> J. Wall,<sup>2</sup> Z. Yang,<sup>5</sup> G. Zane,<sup>2</sup> A. Zhou,<sup>8</sup> and J. Zhou<sup>8</sup>

<sup>1</sup>Lawrence Berkeley National Laboratory, Berkeley, Calif.; <sup>2</sup>University of Missouri, Columbia; <sup>3</sup>Montana State University, Bozeman; <sup>4</sup>University of Washington, Seattle; <sup>5</sup>Oak Ridge National Laboratory, Oak Ridge, Tenn.; <sup>6</sup>Brookhaven National Laboratory, Brookhaven, N.Y.; <sup>7</sup>Pacific Northwest National Laboratory, Richland, Wash.; and <sup>8</sup>University of Oklahoma, Norman

**Project Goals: Model microbial communities can bridge the experimental gap between the simplicity of monocul-**

**ture and the complexity of open environmental systems. We are developing laboratory communities of increasing phylogenetic, functional, and spatial complexity in order to: 1) identify extended genomic and regulatory networks and assess their importance to processes in the field, 2) determine how community organization influences activity and resilience, 3) determine how evolution affects community structure and activity over extended periods of adaptation, as is relevant to extended nutrient stimulation of field sites and the consequences of previous field site manipulation on achieving remediation goals.**

Initial studies examined a relatively simple two-species community in which a sulfate reducer (*Desulfovibrio vulgaris* Hildenborough) was paired with a hydrogen consuming methanogen (*Methanococcus maripaludis*) in suspended steady-state cultures. This two-member community represents a critical step in microbial food webs that control organic matter mineralization. Only when the two species are growing together (syntrophic growth) can organic substrates be oxidized in the absence of electron acceptors such as sulfate. Significant change in community activity was associated with transition from suspended to attached (biofilm) growth states, forming different exopolymers and producing more methane. Methane production was also influenced by the spatial distribution of the two populations, pointing to the importance of initial conditions of biofilm formation for community structure and activity. Although flux balance models were in generally good agreement with experimental data, growth rate and yield of the two-member community increased significantly with long-term maintenance in the lab, indicating that adaptive evolution must be considered in extended manipulation of field sites. In addition, *D. vulgaris* from independently evolving communities had varying competitive abilities when paired with coevolved or foreign *M. maripaludis*, suggesting that there are different adaptive mechanisms and that species interactions affect the course of evolution. Complementary transcriptional and proteome analyses of the two-member community showed that *Desulfovibrio* uses largely independent energy generation pathways during syntrophic growth and sulfate-respiration, and further indicated the importance of using model communities to identify key genetic and physiological processes not expressed in monoculture.

In separate studies, we examined *D. vulgaris* adaptation to salt stress, a common environmental stressor. The most significant improvements in salt tolerance occurred between 500 and 1000 generations in a long-term evolution experiment. Transcriptional and ongoing genome sequencing analyses indicate that the contributing mutations were fixed by 1000 generations, with little subsequent increase in salt tolerance. Ongoing studies are now characterizing the contribution specific mutants identified in the salt-evolved *Desulfovibrio*, and *Desulfovibrio* recovered from evolved two-member communities, to altered phenotypes.

We are also developing higher communities and supporting technologies to understand interactions of increasing complexity, using the two-species community as the foundation for a tri-culture by adding the acetate oxidizing *Geobacter*

*sulfurreducens* as well as tri-cultures of *Clostridium cellulolyticum*, *D. vulgaris*, and *G. sulfurreducens* established in steady state chemostats using cellobiose and different electron acceptors. These higher-complexity communities more closely capture the multiple physiological states of metal-reducing bacteria in the environment, where they are dependent upon associated fermenters for nutrients and reducing equivalents. Parallel studies using steady-state chemostats established from Cr (VII) contaminated groundwater from Hanford, WA simulate the ongoing lactate injection experiments at Hanford, and allow us to better define key populations and interactions that lead to emergent system level properties of nutrient amended subsurface environments. Together these laboratory efforts are designed to improve process control and predictive understanding of environmental microbial activities relevant to DOE priorities.

## 216

### VIMSS Systems Biology Knowledge Base

Paramvir S. Dehal<sup>1,2\*</sup> (PSDehal@lbl.gov), Dylan Chivian,<sup>1,2,3</sup> Adam D. Deutschbauer,<sup>1,2</sup> Jason Baumohl,<sup>1,2</sup> Marcin P. Joachimiak,<sup>1,2</sup> Keith Keller,<sup>1,2</sup> Morgan N. Price,<sup>1,2</sup> and Adam P. Arkin<sup>1,2,3,4</sup>

<sup>1</sup>Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov/>; <sup>2</sup>Lawrence Berkeley National Laboratory, Berkeley, Calif.; <sup>3</sup>DOE Joint BioEnergy Institute, Emeryville, Calif.; and <sup>4</sup>Dept. of Bioengineering, University of California, Berkeley

**Project Goals: The VIMSS Systems Biology Core group seeks to generate and integrate multiple functional genomic data sets in order to create a comprehensive framework for understanding the biology of the sulfate reducing environmental microbe *Desulfovibrio vulgaris* Hildenborough.**

**Background:** The VIMSS Systems Biology Core group seeks to generate and integrate multiple functional genomic data sets in order to create a comprehensive framework for understanding the biology of the sulfate reducing environmental microbe *Desulfovibrio vulgaris* Hildenborough. The group is responsible for high throughput experiments, 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 functional genomics to deal with new data produced by the VIMSS ESPP2 members. The Systems Biology Core is developing methods to store and analyze diverse data sets including: microarrays, ChIP-chip arrays, tiling arrays, proteomics, metabolomics, metabolic flux, phylochips, metagenomics sequencing, genome sequencing, growth curves, phenotype arrays, bar-coded knock out strain collections and links to existing literature and web based resources. Our analysis has been incorporated into our comparative and functional genomics website MicrobesOnline (<http://www.microbesonline.org>) and made available to the wider research community. By

taking advantage of data integration across diverse functional and comparative datasets, we have been able to pursue large research projects in evolutionary and systems biology studies.

**Systems Biology Experiments:** The Systems Biology Core is currently prioritizing the functional annotation of SRB genomes and the creation of a complete systems-level investigation into the physiology of DvH. To accomplish these tasks in the most robust and reproducible manner possible, we have created a robotic setup to automate most of the sample prep, growth curve fitness and phenotype assays and data collection. To functionally annotate SRB genomes, we are systematically generating large sequence-defined transposon libraries in *D. desulfuricans* G20 and DvH. These libraries are “bar-coded” with unique DNA tags which will enable the parallel monitoring of strain fitness in thousands of SRB mutants. By screening across hundreds of growth conditions and monitoring per gene fitness effects, we will be able to assign phenotypic outcomes to each gene.

A complete systems-level investigation into DvH physiology requires a complete parts list of all transcribed elements in the genome. Towards this aim, we are using high-density tiling microarrays and next-generation sequencing technologies to precisely define the transcriptome of DvH. By combining high density tiling arrays with the sequencing data, we will be able to define transcription starts, operon structure, terminator sequences, improve promoter motif predictions, and identify potential antisense transcripts and small RNA genes.

The methods and techniques we have developed are applicable over a broad range of microbes and we will have to capacity to reproduce several of these experiments in related species, allowing the results to be analyzed in the broader context of *Desulfovibrio* evolution.

**Data Integration:** Data management, integration and distribution are critical functions for all large projects. A primary goal of the Systems Biology Core is to capture all experimental data from the ESPP2 investigators, including relevant metadata, raw data and processed data, and to make these data sets available through intuitive queries. Our group has developed Experimental Information and Data Repository (<http://vimss.lbl.gov/EIDR/>) and the MicrobesOnline database to provide this functionality. 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 mutant strains, transposon knockout libraries and protein complex data, in addition to new visualization for assessing existing data sets such as the phenotype microarrays.

**RegTransBase and RegPrecise:** We have built tools and resources for studying regulation in bacteria and archaea using comparative genomics approach. In addition to working on a high quality semi-manual regulon inference in a wide range of species we are building several on-line resources covering different aspects of regulation. RegTransBase, a database of regulatory interactions from literature

collected by a group of experts, currently includes 5100 annotated articles describing 12 thousand experiments. RegPrecise describes manually curated computational predictions of regulons in bacterial genomes done by comparative genomics. RegPredict is a set of highly integrated web tools for fast and accurate inference of regulons. All regulation-related resources are based on the MicrobesOnline data.

**The MicrobesOnline Database:** The MicrobesOnline database (<http://www.microbesonline.org>) currently holds over 1000 microbial genomes and will be updated semi-annually, providing an important comparative and functional genomics resource to the community. New functionality added this year includes the addition of fungal genomes and the framework for adding additional eukaryotic genomes, 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, improved tools to compare multiple microarray expression data across genes and genomes, phylogenetic profile searches using our high quality species tree, and addition of external microarray data from the Many Microbial Microarrays Database for bacteria and Yeast. Additionally we have begun adding metagenomic data to MicrobesOnline.

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 FastHMM and FastBLAST, methods to speed up the most time consuming process of our analysis pipeline, 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. And our FastTree program allows us to create phylogenetic trees for all gene families, even those with over 100,000 members, so that all genes can be studied within an evolutionary framework.

## 217

### Environmental Stress Pathway Project: Study of *Desulfovibrio vulgaris* Hildenborough

**Adam P. Arkin**<sup>1\*</sup> (aparkin@lbl.gov), Edward E. Baidoo,<sup>1</sup> Kelly Bender,<sup>5</sup> Peter I. Benke,<sup>1</sup> Adam Deutschbauer,<sup>1</sup> Matthew Fields,<sup>3</sup> Terry C. Hazen,<sup>1</sup> Zhili He,<sup>4</sup> Dominique C. Joyner,<sup>1</sup> Jay D. Keasling,<sup>1</sup> Kimberly Keller,<sup>2</sup> Eric G. Luning,<sup>1</sup> Aindrila Mukhopadhyay,<sup>1</sup> Lara Rajeev, Jayashree Ray, Judy D. Wall,<sup>2</sup> Grant Zane,<sup>2</sup> Aifen Zhou,<sup>4</sup> and Jizhong Zhou<sup>4</sup>

<sup>1</sup>Lawrence Berkeley National Laboratory, Berkeley, Calif.; <sup>2</sup>University of Missouri, Columbia; <sup>3</sup>Montana State University, Bozeman; <sup>4</sup>University of Oklahoma, Norman; and <sup>5</sup>Southern Illinois University, Carbondale

**Project Goals: The anaerobic, sulfate reducing bacterium *D. vulgaris* Hildenborough, provides an avenue to examine important microbial metal reducing functions in the environment. A systematic understanding of the physiology of such organisms provides invaluable insights into key metabolic and regulatory networks mechanisms and their evolution. The Environmental Stress Pathway Project (ESPP) builds upon the genetic tools, systems biology and analytical methodologies developed for *D. vulgaris* to obtain an in-depth knowledge of the metabolic capabilities, stress response, adaptation and regulatory networks of physiological states and factors that impact its physiology with respect to its environment. Projects outlined in this poster include cells wide studies, transposon library analyses, targeted studies of regulatory factors and signal transduction systems.**

The anaerobic, sulfate reducing bacterium *D. vulgaris* Hildenborough, provides an avenue to examine important microbial metal reducing functions in the environment. A systematic understanding of the physiology of such organisms provides invaluable insights into key metabolic and regulatory networks mechanisms and their evolution. The Environmental Stress Pathway Project (ESPP) builds upon the genetic tools, systems biology and analytical methodologies developed for *D. vulgaris* to obtain an in-depth knowledge of the metabolic capabilities, stress response, adaptation and regulatory networks of physiological states and factors that impact its physiology with respect to its environment. Projects outlined in this poster include cells wide studies, transposon library analyses, targeted studies of regulatory factors and signal transduction systems.

Functional genomics continues to provide a valuable strategy to gain a broad cell wide understanding of *D. vulgaris* physiology. Having applied these measurements to specific stress response assessment, studies are now being focused on understanding *D. vulgaris* biofilms. Both transcript and protein expression profiles demonstrated that biofilm cells have an altered flux of carbon and energy compared to planktonic cells, which may influence metal-interacting capacity and survivability. In addition planktonic and biofilm cells were also exposed to different concentrations of

dissolved oxygen to establish if exposure had any effects on cells and/or biofilm formation. Results from these studies provide insight to better control the growth of sulfate-reducing biofilms.

To improve the functional annotation of *D. vulgaris* and related genomes, large sequence-defined transposon libraries have been generated in both *D. vulgaris* and *D. desulfuricans* G20. These libraries are “barcoded” with unique DNA tags to enable the parallel monitoring of strain fitness in thousands of mutants. Several broad and targeted studies have been conducted using these libraries. Furthermore, a complete systems-level investigation into *D. vulgaris* physiology requires a complete parts list of all transcribed elements in the genome. To address this a high-density tiling microarray and next-generation sequencing technologies are being used to precisely define the transcriptome of *D. vulgaris*.

A key focus of ESPP is to elucidate the structure and evolution of molecular networks in *D. vulgaris*. Towards this aim, a systematic examination of the two component signal transduction pathways is being undertaken. Two component systems trigger responses to a variety of stress and environmental signals. Approximately 70 such systems are annotated in *D. vulgaris*. Using a library of tagged two component system proteins (Histidine Kinases and Response regulators) high throughput protein-DNA interaction strategies are being adopted to map the two component regulatory network. These studies are being complimented using gene deletion mutants in Histidine kinases. The *D. vulgaris* genome also encodes Crp-Fnr like genes that are known to serve as positive transcription factors and play an important role in response to environmental stresses. To examine the function, regulation, and possible networks of these regulators, individual knockout mutants in all four Crp-Fnr genes are being studied. Findings from these studies will be summarized.

## 218

### The Metalloproteomes of Microorganisms are Largely Uncharacterized: A Component of the MAGGIE Project

Michael Thorgersen<sup>1\*</sup> (mthorger@uga.edu), Aleksandar Cvetkovic,<sup>1\*</sup> Angeli Lal Menon,<sup>1</sup> Farris L. Poole II,<sup>1</sup> Joseph Scott,<sup>1</sup> W. Andrew Lancaster,<sup>1</sup> Jeremy Praissman,<sup>1</sup> Sarat Shanmukh,<sup>1</sup> Ewa Kalisiak,<sup>2</sup> Sunia Trauger,<sup>2</sup> Gary Siuzdak,<sup>2</sup> Steven M. Yannone,<sup>3</sup> John A. Tainer,<sup>3</sup> and **Michael W.W. Adams<sup>1</sup>**

<sup>1</sup>Dept. of Biochemistry and Molecular Biology, University of Georgia, Athens; <sup>2</sup>Center for Mass Spectrometry, Scripps Research Institute, La Jolla, Calif.; and

<sup>3</sup>Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.

**Project Goals: The overall goal of the MAGGIE project is to provide robust GTL technologies and comprehensive characterization to efficiently couple gene sequence and**

**genomic analyses with protein interactions and thereby elucidate functional relationships and pathways. The operational principle guiding MAGGIE objectives is that protein functional relationships involve interaction mosaics that self-assemble from independent protein pieces that are tuned by modifications and metabolites, including metals. The objective is therefore to comprehensively characterize protein complexes on a genome-wide basis, including metal-protein complexes, underlying microbial cell biology.**

The overall goal of the MAGGIE project is to provide robust GTL technologies to efficiently couple gene sequence and genomic analyses with protein interactions and thereby elucidate functional relationships and pathways. This involves a comprehensive characterization on a genome-wide scale of protein complexes and protein-cofactor complexes. Metals have been found to play essential roles as cofactors in many enzymes and proteins involved in nearly all cellular processes. However, the true extent of the metalloproteome of any organism remains unknown. It is not possible to predict the types of metal that an organism uses or the number and/or types of metalloprotein encoded in its genome sequence because metal coordination sites are diverse and not easily recognized. In this study, we used *Pyrococcus furiosus*, a hyperthermophilic archaeon that grows optimally at 100°C, as the model organism to gain insight into metalloprotein diversity. The key questions were, what elements does the organism assimilate from its normal laboratory growth medium, can techniques be devised to identify novel metalloproteins, and how specific is metal incorporation into proteins? Conventional non-denaturing liquid chromatography and high-throughput tandem mass spectrometry were used to separate and identify proteins, respectively, and metals were identified by inductively coupled plasma mass spectrometry (ICP-MS). Statistical and algorithmic methods were used to identify potentially novel metalloproteins. A total of 345 metal peaks were identified after fractionation of the cytoplasmic extract of *P. furiosus* through two levels of column chromatography, 160 of which could not be ascribed to any known or predicted metalloprotein. The peaks observed included metals known to be utilized by *P. furiosus* (Fe, Ni, Co, and Zn) and metals that *P. furiosus* was not previously known to take up (U, Pb, Ge, Mo, Mn, and V). Similar chromatographic and metal analyses were performed using the cytoplasmic extracts of two other microorganisms, *Sulfolobus solfataricus* and *Escherichia coli* (grown on their conventional laboratory media). These revealed peaks of yet other types of unanticipated metals. Several of the unassigned metal peaks in the *P. furiosus* analyses were purified to give a single protein using conventional liquid chromatography fractionation. This led to the identification of multiple new types of Ni- and Mo-containing proteins, the properties of which will be presented. The presence of novel V, Mn, Pb, U, and Ge proteins will also be discussed in terms of whether their incorporation into proteins by the organism appears to be intentional or unintentional. These results indicate that metalloproteomes are much more extensive than previously recognized, and likely involve both biologically conventional

and unanticipated metals with implications for a complete understanding of cell biology.

This research was funded by the Department of Energy (DE-FG02-07ER64326) as part of the MAGGIE project.

## 219

### Access to Shape and Assembly of Macromolecular Complexes in Pathways Using Small Angle X-Ray Scattering

Greg L. Hura,<sup>1\*</sup> Michal Hammel,<sup>1</sup> Angeli L. Menon,<sup>2</sup> Robert Rambo,<sup>3</sup> Michael W.W. Adams,<sup>2</sup> and **John A. Tainer**<sup>3,4</sup> (jat@scripps.edu)

<sup>1</sup>Physical Bioscience Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.; <sup>2</sup>Dept. of Biochemistry and Molecular Biology, University of Georgia, Athens; <sup>3</sup>Life Science Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.; and <sup>4</sup>The Scripps Research Institute, La Jolla, Calif.

**Project Goals: The operational principle guiding MAGGIE (Molecular Assemblies, Genes, and Genomics Integrated Efficiently) objectives can be succinctly stated: protein functional relationships involve interaction mosaics that self-assemble from independent macromolecular pieces that are tuned by modifications and metabolites. Several metrics and tools have been developed in MAGGIE which identify, capture, characterize and predict the effects of macromolecular assemblies. One of our developed tools is small angle X-ray scattering (SAXS). MAGGIE was the first to recognize and apply SAXS as a proteomic scale tool by coupling robotic fluid handling with the extreme brightness of synchrotron light.**

The operational principle guiding MAGGIE (Molecular Assemblies, Genes, and Genomics Integrated Efficiently) objectives can be succinctly stated: protein functional relationships involve interaction mosaics that self-assemble from independent macromolecular pieces that are tuned by modifications and metabolites. Several metrics and tools have been developed in MAGGIE which identify, capture, characterize and predict the effects of macromolecular assemblies. One of our developed tools is small angle X-ray scattering (SAXS). MAGGIE was the first to recognize and apply SAXS as a proteomic scale tool by coupling robotic fluid handling with the extreme brightness of synchrotron light.

SAXS characterizes the shape and assembly of macromolecules in solution. Information both on routine cellular processes and cell state are often carried in the shape and assembly of macromolecules on the length scales measured by SAXS. As we have enabled data collection at a rate of 96 samples in 4 hours on practical concentrations and volumes for most macromolecules a new scale of structural characterization has been enabled. Facile access to shape and conformation of macromolecules in a large number of

contexts which are encountered during cellular processes has had significant impact on our understanding of pathways and mechanisms for information transduction. Often macromolecules interact with several metabolites and specific combinations cause large length scale re-arrangements.

Pathways which maintain genomic stability, protein re-naturation and most recently components of energy generation have been targets of our approach. Our purification and data collection success rate has been significantly enhanced using proteins from our 3 target extremophile organisms. In addition to providing higher stability proteins and complexes we hope to learn the mechanisms these organisms utilize to accomplish these ubiquitous tasks under the challenging conditions in which they thrive. These may provide design principles for engineering pathways of interest which are of ultimate interest to GTL.

In addition to utilizing SAXS as a tool to characterize shape and assembly of pathway components, members may be identified which are tractable for other biophysical techniques. Multimerization state influences concentrations required for many techniques. Large flexible regions may prohibit crystallization. Specific alterations to sequence or post translational modifications may stabilize a conformation which enhances the likelihood of success by other techniques.

The developed infrastructure for SAXS is a unique resource for GTL and has already been utilized by hundreds of research labs around the country. Future target pathways and relevant components will be directed by the bioinformatics core of MAGGIE. In addition, the extension of SAXS to membrane proteins is currently being explored.

### Molecular Interactions, Protein Complexes, and Structural Biology

## 220

### The ENIGMA Project: Mapping Protein Assemblies and Modifications by Cellular Deconstruction and Mass Spectrometry in the Hyperthermophiles *Sulfolobus solfataricus*, *Pyrococcus furiosus*, and *Halobacterium NRC-1*

Robert Rambo,<sup>1</sup> Adam Barnebey,<sup>1</sup> Michael W. Adams,<sup>2</sup> Gary Siuzdak,<sup>3</sup> Sunia Trauger,<sup>3</sup> Nitin S. Baliga,<sup>4</sup> Stephen R. Holbrook,<sup>1</sup> Trent Northen,<sup>1</sup> Ben Bowen,<sup>1</sup> Greg Hura,<sup>1</sup> John A. Tainer,<sup>1,5</sup> and **Steven M. Yannone**<sup>1\*</sup> (SMYannone@lbl.gov)

<sup>1</sup>Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.; <sup>2</sup>Dept. of Biochemistry and Molecular Biology, University of Georgia, Athens; <sup>3</sup>Center for Mass Spectrometry, Scripps Research Institute, La

Jolla, Calif.; <sup>4</sup>Institute for Systems Biology, Seattle, Wash.; and <sup>5</sup>Dept. of Biochemistry and Molecular Biology, Scripps Research Institute, La Jolla, Calif.

**Project Goals:** As part of the integrated ENIGMA project our goals are to develop generalized methodologies for identifying and isolating molecular machines and modified proteins from native biomass. We aim to identify and characterize both stable and transient macromolecular assemblies within the cell, the cell membrane, and outside of the cell using a rational cellular deconstruction approach. Proteins are identified by MS/MS from our cellular deconstruction fractions have inherent cellular locality, assembly mass, protein modification, and potential partner protein information associated with their identification. Our proteome-wide approach to identify macromolecular assemblies and protein modifications will ultimately lead to identification of metabolic modules suitable for transfer between microbes. We are developing generally applicable molecular and biophysical technologies for GTL to ultimately apply functional modules to confer specific metabolic capabilities to microbes and address DOE mission goals.

Dynamic protein-protein interactions are fundamental to most biological processes and essential for maintaining homeostasis within all living organisms. These interactions create dynamic and diverse functional networks essential to biological processes. Thus, a thorough understanding of these networks will be critical to engineering biological processes for DOE missions. The 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 the ENIGMA project, we have developed methods for whole cell deconstruction of microbes that separates intact protein complexes under native conditions. The cellular deconstruction primarily fractionates the organism into four major classes; 1) membrane, 2) large mass (>800 kDa), 3) small mass (<800 kDa), and 4) extracellular. The final cellular partitions were resolved on SDS-PAGE and excised for high throughput MS/MS protein identification at the Scripps Center for Mass Spectrometry. We have achieved over 50 percent coverage of the predicted proteome and all identifications have inherent subcellular locality and co-fractionation information associated with them: 1) proteins identified in sedimenting fractions are by definition members of large mass complexes, 2) isolated membrane vesicles allow identification of membrane associated proteins, and 3) extracellular proteins contain extracellular membrane-bound and secreted proteins groups. We are refining analysis tools to identify co-variations of proteins across the fractionation scheme to build a system-wide protein interaction network. Ultimately, we aim to identify metabolic modules suitable to transfer specific metabolic processes between microbes to address specific DOE missions while develop-

ing generally applicable molecular and biophysical technologies for GTL.

## 221

### High-Pressure Cryocooling of Protein Crystals: Applications to Understanding Pressure Effects on Proteins

**Sol M. Gruner**\*<sup>1,2</sup> (smg26@cornell.edu), Chae Un Kim,<sup>3</sup> Nozomi Ando,<sup>1</sup> Buz Barstow,<sup>1</sup> Mark Tate,<sup>1</sup> and Yi-Fan Chen<sup>3</sup>

<sup>1</sup>Physics Dept., <sup>2</sup>Cornell High Energy Synchrotron Source (CHESS), and <sup>3</sup>MacCHESS (Macromolecular facility at the Cornell High Energy Synchrotron Source), Cornell University, Ithaca, New York

<http://bigbro.biophys.cornell.edu>

**Project Goals:** Understand effects of pressure on proteins.

A novel high-pressure cryocooling technique for preparation biological samples for x-ray analysis has been developed [1-3]. The method, high-pressure cryocooling, involves cooling samples to cryogenic temperatures (~ 100 K) in high-pressure Helium gas (up to 200 MPa). It bears both similarities and differences to high-pressure cooling methods that have been used to prepare samples for electron microscopy, and has been especially useful for cryocooling of macromolecular crystals for x-ray diffraction. Many different kinds of macromolecular crystals have been successfully high-pressure cryocooled and excellent crystal diffraction has been obtained with little or no penetrating cryoprotectants.

This new method has great potential for understanding pressure effects on proteins. As an example, high pressure cryocooling has been used to understand the structural basis for why the emission spectrum of the protein, Citrine, is pressure dependent [4,5]. The deformation of the Citrine chromophore is actuated by the differential motion of two clusters of atoms that compose the  $\beta$ -barrel scaffold of the molecule, resulting in a slight bending of the  $\beta$ -barrel. The high-pressure structures also reveal a perturbation of the hydrogen bonding network that stabilizes the excited state of the Citrine chromophore. The perturbation of this network is implicated in the reduction of fluorescence intensity of Citrine. The blue-shift of the Citrine fluorescence spectrum resulting from the bending of the  $\beta$ -barrel provides structural insight into the transient blue-shifting of isolated yellow fluorescent protein molecules under ambient conditions and suggests mechanisms to alter the time-dependent behavior of Citrine under ambient conditions. Finally, the Citrine example serves as a model for the way in which global pressure-induced structural perturbations affect the activity of proteins, as well as a model for how these perturbations may be studied.

Supported by the Dept. of Energy (DE-FG02-97ER62443), the National Institutes of Health (RR-001646) and CHESS

(National Science Foundation and NIH-NIGMS through NSF DMR-0225180.)

## References

1. Kim et al., Proc. Natl. Acad. Sci., USA 106 (2009) 4596.
2. Kim et al., J. Appl. Cryst., 41 (2008) 1.
3. Kim et al., Acta Cryst. D63 (2007) 653.
4. Barstow et al., Biophys. J., 97 (2009) 1719.
5. Barstow et al., Proc. Natl. Acad. Sci. USA, 105 (2008) 13362.

# 222

## Opportunities for Structural Biology and Imaging at NSLS-II

Lisa M. Miller<sup>1,2\*</sup> (lmliller@bnl.gov) and Wayne A. Hendrickson<sup>1,3</sup>

<sup>1</sup>NSLS-II Project and <sup>2</sup>National Synchrotron Light Source, Brookhaven National Laboratory, Upton, N.Y.; and <sup>3</sup>Dept. of Biochemistry and Molecular Biophysics, Columbia University, New York, N.Y.

<http://www.bnl.gov/nsls2/>

**Project Goals: In this poster, we will describe how the unique characteristics of NSLS-II can address scientific problems relevant to BER scientists by presenting a series of applications to bioenergy, carbon cycling and sequestration, and contaminant transport and cleanup in the environment. We will emphasize the wide range of techniques that will permit multiscale exploration: at the molecular level, to understand how genes determine biological structure and function; at the cellular level, to understand how molecular processes are coordinated to execute cell function; and at the level of microbial communities and higher organisms to understand how cells interact and respond to their environment. Finally, we are eager to solicit new applications of synchrotron science to research problems of interest to BER scientists.**

The National Synchrotron Light Source-II (NSLS-II), which is now under construction at Brookhaven National Laboratory, will provide a broadband source of synchrotron photons from infrared light to x-rays with a brightness unsurpassed by any synchrotron facility worldwide. This new facility is scheduled to be operational in 2015 and will replace the existing NSLS.

The extreme brightness and coherence of NSLS-II will enable characterization techniques, such as nanoscale imaging, that are currently in their infancy or do not exist today. But importantly, NSLS-II will also take widely utilized methods, such as macromolecular crystallography (MX), small-angle x-ray scattering (SAXS), and x-ray absorption spectroscopy (XAS), and extend them to new regimes in time- and spatial-resolution that cannot be achieved today.

Recent workshops by BER have identified both biological challenges and technological needs that are important to the BER research community. For example, in May 2009, BER held the New Frontiers in Characterizing Biological Systems workshop to address the next generation challenges in genomics science and its connection to functional systems. The panel identified numerous knowledge gaps that inhibit the understanding of biological systems. These knowledge gaps are relevant to understanding research areas paramount to BER interest, including the generation and processing of biomass into chemical energy, climate change and the cycling of carbon and nutrients, and the transformation of natural and man-made contaminants in the environment.



The NSLS-II will be operational in 2015 and will provide a broadband source of photons with a brightness and coherence unsurpassed by any synchrotron worldwide. For BER science, it will enhance time-resolved structural studies of macromolecules and complexes, especially in more natural environmental settings. High throughput structure/function determination will be optimized to link genomic information to molecular events. And it will provide a wide range of nanoscale imaging capabilities with the possibility of multi-modality characterization of identical samples.

Synchrotron-based characterization tools are well-suited to fill the identified gaps. Synchrotron studies will generate basic understanding of biological processes, and not just for particular phenomena at a certain physical or temporal scale, but as linked pan-genomically across scales of investigation. With the high brightness and coherence of NSLS-II, structural studies of macromolecules and complexes will be possible in a time-resolved manner, especially in more natural environmental settings. Moreover, high throughput structure/function determination will be able to link genomic information to molecular events. NSLS-II will provide a wide range of nanoscale imaging capabilities, permitting multi-modality characterization of identical samples.

At the NSLS today, biological and environmental sciences users represent approximately 60% of the user community and more than 650 of the facility's annual publications. NSLS-II plans to follow in the footsteps of the current NSLS by providing a wide range of characterization techniques to the biological sciences community. In January 2008, a series of Scientific Strategic Planning workshops were held at the NSLS to identify a pathway forward to NSLS-II. An overarching conclusion from the Life and Environmental Sciences workshops was the desire within these communities to see increased interaction, collaboration, multi-technique integration, and cross-disciplinary approaches to doing science in the future. It was suggested that this mode of research can be achieved through a "Biol-

ogy Village” environment, which would include strategically locating beamlines for scientific interaction, having programmatic overlap through shared equipment, technology, and human resources, and playing an active role in the Joint Photon Sciences Institute (JPSI), an interdisciplinary facility at BNL that will facilitate R&D efforts.

In this poster, we will describe how the unique characteristics of NSLS-II can address scientific problems relevant to BER scientists by presenting a series of applications to bioenergy, carbon cycling and sequestration, and contaminant transport and cleanup in the environment. We will emphasize the wide range of techniques that will permit multiscale exploration: at the molecular level, to understand how genes determine biological structure and function; at the cellular level, to understand how molecular processes are coordinated to execute cell function; and at the level of microbial communities and higher organisms to understand how cells interact and respond to their environment. Finally, we are eager to solicit new applications of synchrotron science to research problems of interest to BER scientists.

## Validation of Genome Sequence Annotation

# 223

### Robotic Chemical Protein Synthesis for the Experimental Validation of the Functional Annotation of Microbial Genomes

**Stephen Kent** and Kalyaneswar Mandal\* (kmandal@uchicago.edu)

Institute for Biophysical Dynamics, University of Chicago, Ill.

**Project Goals: Robotic total chemical synthesis to make proteins and protein domains, for the validation of functional annotation of predicted open reading frames.**

Modern total protein synthesis has evolved from the ‘chemical ligation’ methods introduced by the Kent laboratory [Kent SBH. Total chemical synthesis of proteins. *Chemical Society Reviews* 2009; 38: 338-51.]. Unprotected synthetic peptide segments, spanning the amino acid sequence of the mature polypeptide chain derived from a predicted open reading frame, are covalently joined to one another by chemo-selective reaction. Native chemical ligation, the thioester-mediated covalent bond-forming chemoselective reaction of unprotected peptides at a Cys residue, is the most robust and useful ligation chemistry developed to date. The synthetic protein is then used to experimentally validate the predicted biochemical function, and in selected cases to determine the Xray structure of the protein molecule (Figure).

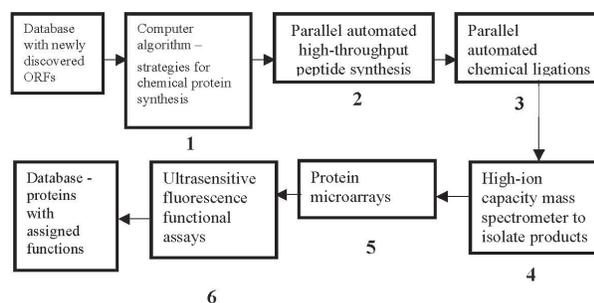


Figure 1. Modular high-throughput platform for fast and parallel total chemical synthesis, mass-spectrometric purification and single-molecule spectroscopic assay to annotate function for newly predicted proteins.

We have previously described the use of x,y,z robotics and laboratory automation and efficient Fmoc chemistry SPPS protocols for the simultaneous parallel synthesis of the key peptide-thioester building blocks needed for chemical protein synthesis. This made use of a recently reported novel resin linker [Blanco-Canosa JB, Dawson PE: An efficient Fmoc-SPPS approach for the generation of thioester peptide precursors for use in native chemical ligation. *Angew Chem Int Ed Engl.* 2008, 47:6851-5]. Typical data are shown in Figure 2 (Top).

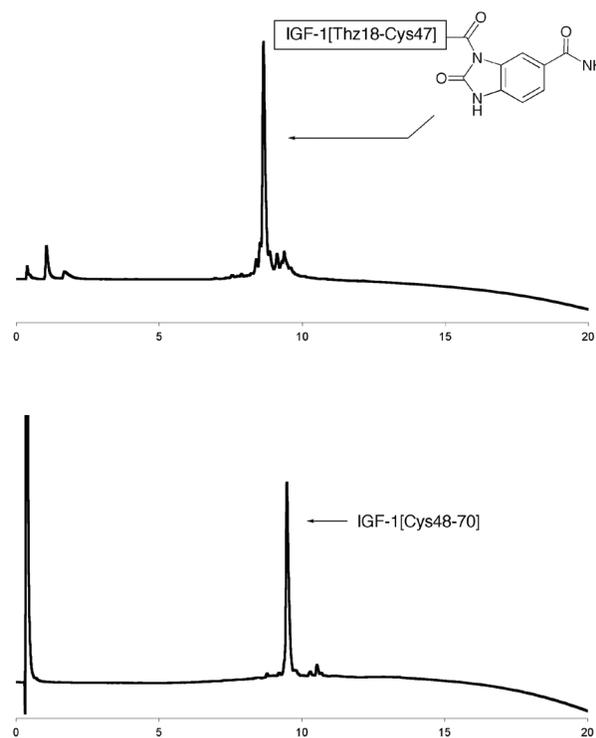


Figure 2. Automated robotic Fmoc SPPS preparation of peptide-thioesters. HPLC-electrospray MS of crude products are shown.

Ready preparation of peptide-thioesters enables the straightforward total chemical synthesis of proteins by native chemical ligation. Proof-of-concept total chemical

syntheses of predicted proteins from microbial and plant genomes will be presented.

## 224

### Using Deep RNA Sequencing for the Structural Annotation of the *Laccaria bicolor* Mycorrhizal Transcriptome

Peter E. Larsen,<sup>1\*</sup> Geetika Trivedi,<sup>2</sup> Avinash Sreedasyam,<sup>2</sup> Vincent Lu,<sup>1</sup> Gopi K. Podila,<sup>2</sup> and Frank R. Collart<sup>1</sup> (fcollart@anl.gov)

<sup>1</sup>Biosciences Division, Argonne National Laboratory, Lemont, Ill. and <sup>2</sup>Dept. of Biological Sciences, University of Alabama, Huntsville

[http://www.bio.anl.gov/molecular\\_and\\_systems\\_biology/proteins.html](http://www.bio.anl.gov/molecular_and_systems_biology/proteins.html)

**Project Goals: To facilitate the process for acquisition of function from complex environmental sequence data sets, we developed methods to utilize RNA-seq data to validate current gene model intron-exon boundary, correct errors in the structural annotation and extend the boundaries of the current gene models using assembly approaches.**

Advances in sequences technology have enabled deep sequence interrogation of individual organisms as well as complex systems. This capability has led to an improved appreciation of the biological diversity associated with specific ecosystems and the complexity of the molecular systems involved in perception and response to external stimuli. Mapping these signaling pathways is challenging however in sequence data sets from environmental and/or metagenomic projects where uncharacterized organisms often represent a high proportion of the sequence data. To facilitate the process for acquisition of function from complex environmental sequence data sets, we are developing methods to utilize RNA-seq data to correct errors in the structural annotation and extend the boundaries of current gene models using assembly approaches. To validate the methods, we used a transcriptomic data set derived from the fungus *Laccaria bicolor* which develops a mycorrhizal symbiotic association with the roots of many tree species, in which the fungus provides nutrients to the tree in exchange for photosynthetically-derived sugars. This fungal-plant symbiosis is a widespread process of major ecological importance and knowledge of the molecular events and expressed protein sequences associated with the development of the mycorrhizal system is essential for our understanding of natural biological processes related to carbon sequestration, carbon management, sustainability and bioenergy.

We generated >30 million RNA-seq reads from *Laccaria* grown in culture. Our study used the 20614 gene “best model” set and 65-megabase *Laccaria* genomic DNA sequence from the publically available FTP site at the Joint Genome Institute. Our analysis focused on the subset of 1501 gene models that are differentially expressed in the mycorrhizal transcriptome and are expected to be important

elements related to carbon metabolism, membrane permeability and transport, and intracellular signaling.

Our analysis of the intron-exon boundaries in current JGI best gene model set indicates the quality of *L. bicolor* structural annotation is enabling for homology-based comparison applications, but has severe limitations for experimental studies. For every intron-exon boundary in JGI Best Model set, we generated an 18-mer ‘probe’ sequence consisting of 9 bp up and down-stream of the intron-exon boundary. This intron-spanning sequence was used to search the set of RNA-seq reads. At least one read-containing probe was considered validation of gene model intron-exon boundary. Using these criteria, we were able to validate ~80% of the intron-exon boundaries within the gene model boundaries. This level of validation is notable in view of the complexity of the fungal genome (*L. bicolor* genes contain an average of 5.4 introns) and the annotation limitations arising from the relatively small number of sequenced fungal genomes. However, the combination of the error rate and intron density means that 42% of the current gene models contain intron/exon boundaries that do not map to the mRNA sequence data. Also, 58% of gene model 5’ and/or 3’ boundaries did not agree with the collected transcriptomic data. Inaccurate representations of the protein coding sequence are a consequence of these inconsistencies. Accurate coding sequences are essential for experimental approaches to characterize protein function and also to enhance the utility of tools that enable identification cellular localization signals and functional domains. Substantial changes to predicted UTRs also affect the ability to predict the regulatory mechanisms of mycorrhizae-specific genes. To improve the experimental utility of the gene model set, we developed algorithms that use the RNA-seq data to extend the boundaries of the current gene model set where appropriate, identify those intron-exon boundaries that can be validated by the transcriptomic data, and to generate novel intron-exon boundaries to bridge those regions of the gene models that are not supported by RNA-seq data. This extended and bridged contiguous expressed sequence was then aligned to the genome using a modified Smith-Waterman algorithm to recover gene model’s structural annotation. Of the set of 1501 gene models, 1439 (96%) successfully generated modified gene models in which all error flags were successfully resolved and sequences aligned to genomic sequence. The remaining 4% (62 gene models) either had deviations from transcriptome data that could not be spanned or generated sequence that did not align to genomic sequence. We considered a gene model significantly changed if at least one of three criteria were met: 1) an inconsistency in the original gene model was successfully bridged and aligned to scaffold, 2) the revised gene model contained a change in the total number of exons, and/or 3) we observed an absolute change in expressed gene size of more than 10%. Based on application of these criteria to the set of 1439 revised models, 974 (69%) of gene models required changes to match the transcriptomic consensus sequence. Additionally, for 465 (31%) of the models in the original best gene model set, we did not detect any inconsistencies and therefore have independently confirmed the previously published ‘BestModel’ annotation. Of those 62 gene models that could not be adequately

validated by the method proposed here, a number appear to have multiple isoforms in the expressed transcriptome data identifying them as genes of potential biological interest.

The outcome of this process is a set of high confidence gene models that can be reliably used for experimental characterization of protein function. This improved annotation process can be extended to other important gene families and will facilitate the process to identify the molecular mechanisms leading to the development of the mycorrhizal symbiosis and its implications in improving carbon sequestration by poplar.

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357.

## 225

### Molecular Approaches for Elucidation of Sensory and Response Pathways in Cells

Sarah Zerbs,<sup>1\*</sup> Sarah E. Giuliani,<sup>1</sup> Elizabeth Landorf,<sup>1</sup> Maureen McNulty,<sup>1</sup> William Studier,<sup>2</sup> and Frank R. Collart<sup>1</sup> (fcollart@anl.gov)

<sup>1</sup>Biosciences Division, Argonne National Laboratory, Lemont, Ill. and <sup>2</sup>Biology Division, Brookhaven National Laboratory, Upton, N.Y.

[http://www.bio.anl.gov/molecular\\_and\\_systems\\_biology/proteins.html](http://www.bio.anl.gov/molecular_and_systems_biology/proteins.html)

**Project Goals:** This program addresses the hypothesis that cellular behavior can be modeled through an understanding of the biological interface with the environment and the cellular responses that originate from the cell/environment interaction. The long term objective of the program is to define cellular sensory and regulatory pathways that respond to environmental nutrients thereby facilitating a system-level model that predicts the cellular response to environmental conditions or changes.

Increased knowledge of protein function enhances our understanding of cellular functions and is ultimately required to model biological activities and systems. This program addresses the hypothesis that cellular behavior can be modeled through an understanding of the biological interface with the environment and the cellular responses that originate from the cell/environment interaction. The long term objective of the program is to define cellular sensory and regulatory pathways that respond to environmental nutrients thereby facilitating a system-level model that predicts the cellular response to environmental conditions or changes. The program uses a parallel strategy of technology development to improve capabilities for extraction of relevant biological information from the sequence data coupled to genome scale approaches for elucidation of protein function and cellular regulatory networks.

One aspect of this program will develop tools to bridge the gap between genomes and systems biology. Progress in sequencing technology has provided molecular validation of the diversity and complexity of environmental systems. However, sequencing capacity has far outpaced computational and experimental methods to fully utilize the genomic data. We are addressing this gap between DNA sequence and the ability to extract relevant biological information from the sequence data by the development of genome scale approaches for elucidation of protein function and cellular regulatory networks. These approaches utilize next generation sequencing technology and high throughput approaches to enable economical and efficient protein production and characterization.

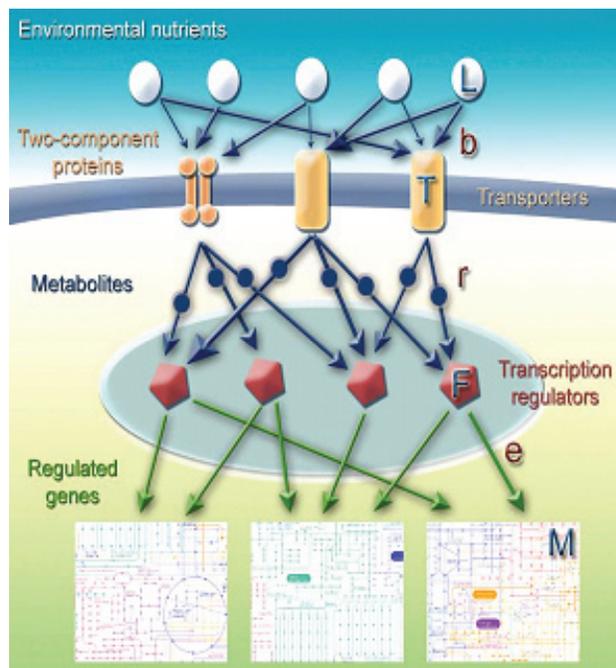


Fig 1. Illustration of experimental approach and application to systems modeling.

The capabilities to improve functional interrogation of sequences are coupled to *in vitro* methods for functional characterization of proteins involved in cellular sensory and response pathways. The functional screens will focus on key proteins that mediate communication between the cell and the environment such as transporters, two-component sensory systems, and membrane receptors (Fig. 1). This functional characterization will be linked to the cellular regulatory network by identification of the transcription factors whose activity is mediated by the environmental ligands or their metabolic derivatives. A coupling of the regulatory ligands with the DNA-binding regions of the transcription factors allows the association of metabolic pathways with the regulatory network. This genome scale process will determine the functional properties and potential of microbes and plants that are central to DOE missions. The functional assignments and ability to define specific sensory and regulatory pathways will increase the predictive capability of current models and support the development of

predictive systems-level models. This increased knowledge of the molecular components and control features of cellular sensory and response pathways is essential for our understanding of natural biological processes related to carbon management, sustainability and bioenergy.

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory (“Argonne”). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. Brookhaven National Laboratory is operated under Contract No. DE-AC02-98CH10886.

## 226

### Functional Linkage of ABC Transporter Profile with Metabolic Capability in *Rhodopseudomonas palustris*

Sarah E. Giuliani,<sup>1\*</sup> Ashley M. Frank,<sup>1</sup> Catherine Seifert,<sup>1</sup> Lisa M. Miller,<sup>2</sup> Loren Hauser,<sup>3</sup> and Frank R. Collart<sup>1</sup> (fcollart@anl.gov)

<sup>1</sup>Biosciences Division, Argonne National Laboratory, Lemont, Ill.; <sup>2</sup>National Synchrotron Light Source, Brookhaven National Laboratory, Upton, N.Y.; and <sup>3</sup>Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

[http://www.bio.anl.gov/molecular\\_and\\_systems\\_biology/proteins.html](http://www.bio.anl.gov/molecular_and_systems_biology/proteins.html)

**Project Goals: We suggest that the functional profile of the genome set of transporter proteins is predictive of metabolic capabilities and ecological niche of organisms. To test this hypothesis, we profiled the genome set of ABC transporters for *Rhodopseudomonas palustris* CGA009 to define the relationship between the transporter profile and metabolic capability for *R. palustris* CGA009.**

Transporter proteins are an organism’s primary interface with the environment. The expressed set of transporters mediates cellular metabolic capabilities and influences signal transduction pathways and regulatory networks. The role and impact of different transporters families differ in eukaryotic and prokaryotic organisms and the absolute number of transporters is dependent on the characteristics of the ecological niche. We suggest that the functional profile of the genome set of transporter proteins is predictive of metabolic capabilities and ecological niche of organisms. To test this hypothesis, we profiled the genome set of ABC transporters for *Rhodopseudomonas palustris* CGA009. In the *R. palustris*, ABC-type transporters represent approximately 45% of all transporters encoded in the genome. The ABC transporters family is widely distributed in soil organisms and can transport a variety of substrates such as metals, small ions, mono- and oligosaccharides, peptides, amino acids, iron-siderophores, polyamines, and vitamins.

An ABC transporter complex consists of a permease, ATPase, and a solute binding protein. The ligand specificity is determined by the solute binding protein which in

some cases can utilize multiple membrane permeases. The genome of *R. palustris* CGA009 encodes approximately 117 ABC type transporters as determined by the number of encoded solute binding proteins. The functional properties of these transport proteins are largely unknown and less than 10% have specific functional assignments. The largest group of binding proteins is annotated as “branched-chain amino acid” binding protein. To improve the utility of the function annotation, we expressed and purified the set of binding proteins from *R. palustris* and are characterizing ligand-binding specificity using ligand libraries consisting of environmental and cellular metabolic compounds and high throughput binding screens, including fluorescence thermal shift, small angle x-ray scattering, x-ray absorption spectroscopy, circular dichroism spectroscopy, and infrared spectroscopy. To date, this process resulted in the assignment of specific binding ligands for approximately 60% of the purified and screened proteins. In most cases, the binding was observed for specific compound classes and was observed for only 1-3 compounds from the entire ligand library. For approximately 20% of the screened proteins, a specific binding ligand was not observed, which we attribute to the limited scope of the screening library relative to the complexity of compounds in the natural environment.

The impact of these studies is two-fold. First, our screening method generated specific functional annotations for an important group of uncharacterized or incorrectly annotated transporter proteins. For example, six proteins encoded by genes annotated as branched chain amino acid binding proteins were demonstrated to bind various aromatic compounds derived from lignin degradation. Analysis of the flanking genomic regions reveals the co-localization of these transporter genes with metabolic genes associated with utilization of the transported compounds. Similar functional insight was obtained for previously uncharacterized proteins associated with the transport of fatty acids, dicarboxylic acids, oligopeptides, metals, and additional small molecule compounds. This functional insight can be used to improve the annotation of related organisms and provides a route to evaluate the evolution of the important and diverse group of transporter proteins.

Second, the results of this study also provide important biological insight for the metabolic capabilities and environment fitness of this organism. The profile and number of transport proteins specific for aromatic compounds is consistent with ecological and laboratory studies which demonstrate the capabilities of this organism for the utilization of plant degradation products such as lignin-derived aromatic compounds.

One of these binding proteins, RPA1385, showed high affinity and selectivity for vanadate, which is a catalytic component of a nitrogenase protein complex. *R. palustris* is a nitrogen fixing bacteria and has been shown to utilize a vanadium nitrogenase (V-nitrogenase) as a metabolic alternative when molybdenum is limited in the environment. Prior to this research, the cyanobacterium *Anabaena variabilis* (which also contains a V-nitrogenase) was the only organism known to contain a defined high-affinity vanadate

transport system. In *R. palustris*, genes RPA1381-1386 are annotated as components of a vanadate nitrogen fixation system based on homology to other similar proteins. However, in *R. palustris*, homology search approaches failed to identify the high-affinity vanadate transport system. Our ligand mapping approach identifies the RPA1385 protein as the vanadate SBP gene for this ABC transport system. This finding not only identifies a key component of the vanadate nitrogenase fixation pathway for this organism, but may also confirm a proposed hypothesis that the presence of this system in *R. palustris* suggests vanadate transport systems have evolved at least twice from dissimilar ancestral genes.

The functional assignments in conjunction with gene expression profiles and transcription factor DNA binding sites enable the identification of the cellular regulatory and metabolic components that enable the use of lignin degradation products for cell growth. This approach is being applied to other sequenced strains of *R. palustris* to provide evolutionary insight for the number and substrate specificity of this family of ABC type transporters. These capabilities will enable the identification and characterization of metabolic and regulatory pathways that are associated with a specific environmental niche.

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. Brookhaven National Laboratory is operated under Contract No. DE-AC02-98CH10886.

## 227

### Phylogenomics-Guided Validation of Function for Conserved Unknown Genes

Valérie de Crécy-Lagard<sup>1\*</sup> (vcrecy@ufl.edu), Basma El Yacoubi,<sup>1</sup> Marc Bailly,<sup>1</sup> Ian K. Blaby,<sup>1</sup> Anne Pribat,<sup>2</sup> Aurora Lara-Núñez,<sup>2</sup> and Andrew D. Hanson<sup>2</sup>

<sup>1</sup>Dept. of Microbiology and Cell Science and <sup>2</sup>Dept. of Horticultural Sciences, University of Florida, Gainesville

**Project Goals: Our overall goal is to establish an innovative 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. By implementing this approach we will predict, and experi-**

**mentally validate for a chosen subset, the function of ~1500 unknown genes.**

Identifying the function of every gene in all sequenced organisms is a central challenge of the post-genomic era. We are submerged in genomic, transcriptomic, and proteomic data but the functions of about half (range 20 to 60%) of the genes in any given organism are still unknown. Our goal is to predict and experimentally verify the *in-vivo* function of proteins that lack homologs of known function ('unknown' protein families) and that are highly conserved between prokaryotes and plants. Our approach combines the extensive post-genomic resources of the plant field with the use of comparative genomic tools made possible by the availability of thousands of sequenced microbial genomes. This is an integrative approach to predict gene function whose early phase is computer-assisted, and whose later phases incorporate intellectual input from expert plant and microbial biochemists. It allows bridging of the gap between automated homology-driven annotations and the classical gene discovery efforts driven mainly by experimentalists. Our goal is to predict and experimentally validated the function of 15 "unknown" protein families". We have already validated predictions for seven families (in orange in Table 1 and we present the other eight most mature predictions that currently being tested (in yellow in Table 1). Two examples of this second list will be presented in more detail to emphasize the synergistic aspects of plant-microbe comparative genomics.

COG0799 proteins occur in plants, in nearly all bacteria, and in animals and fungi. Plants have two isoforms, one apparently chloroplastic, the other mitochondrial. The archetypal member of the family is the plant Iojap protein; *iojap* mutants of maize lack functional chloroplast ribosomes. In bacteria, COG0799 genes cluster strongly with the NAD synthesis gene *nadD* (nicotinate mononucleotide adenyltransferase) and sometimes the two genes are fused. COG0799 genes also cluster with genes encoding the ribosomal biogenesis protein ObgE and ribosomal proteins L21 and L27, making a connection with the ribosome lesion in the maize *iojap* mutant. Furthermore, transcriptomic data from *Arabidopsis* show co-expression of *iojap* with various chloroplast ribosome protein genes. NadD mediates a reaction in the *de novo* synthesis of NAD and potentially in salvage of nicotinamide mononucleotide (NMN). We therefore predict that Iojap catalyzes a process in ribosome biogenesis that releases NMN from NAD, and that the NMN is recycled by NadD. Possible Iojap reactions include a NAD-dependent DNA ligase-like reaction or an ADP-ribosyltransferase.

Case	Hypothesis	TAIR ID	COG, gene name	Subsystem in SEED	Experimental verification status	PubMed ID
1	Pterin carbinolamine dehydratase with role in Moco metabolism	At1g29810 At5g51110	COG2154, phhB	Pterin_carbinolamine_dehydratase	Validated in in 7 eukaryotes and 8 prokaryotes	18245455
2	t6A biosynthesis	At5g60590	COG0009, YrdC	YrdC-YciO	Validated in Yeast, Archaea and two bacteria;	19287007
3	PTPS family protein replacing the FolB step in folate synthesis	-	COG0720	Experimental-PTPS	Validated in 1 eukaryote and 8 prokaryotes	19395485, 18805734
4	Metal chaperone-Zinc homeostasis	At1g15730, At1g26520, At1g80480	COG0523	COG0523	Validated in several bacteria	19822009
5	Folate-dependent Fe/S cluster synthesis or repair protein	At4g12130 At1g60990	COG0354, ygfZ	YgfZ-Fe-S	Validated in <i>E. coli</i> , <i>Haloferax volcanii</i> , <i>Arabidopsis</i> , <i>Leishmania</i> , yeast, mouse	Submitted
6	Alternative route for 5-formyltetrahydrofolate disposal	At2g20830	COG3643	Experimental_Histidine_Degradation	Verified in 5 prokaryotes	Manuscript in prep
7	t6A biosynthesis	At2g45270, At4g22720	COG0533, YgiD	YrdC-YciO	Validated in yeast	Manuscript in prep
8	NAD-dependent nucleic acid AMP ligase	At3g12930, At1g67620	COG0799, alr4169	Iojap	In progress <i>E. coli</i>	
9	5-Formyltetrahydrofolate cycloligase paralog	At1g76730	COG0212	5-FCL-like_protein	Predicted role in thiamine recycling	
10	Hydroxyproline-galactosyl hydrolase	At5g12950, At5g12960	COG3533, SAV1144	COG3533	In progress in <i>X. campestris</i>	
11	m6A in small rRNA	At4g28830	COG2263	rRNA_modification_Archaea	Mutant analysis in <i>H. volcanii</i> in progress	
12	Choline transporter	NiaP homolog At1g13050	MFS superfamily	Choline transport and metabolism	In progress <i>R. solanacearum</i> and <i>B. xenovorans</i>	
13	Ribosome assembly/translation termination	At1g09150	COG2016	rRNA_modification_Archaea	In progress in yeast and <i>H. volcanii</i>	
14	Phytol phosphate kinase	At1g78620	COG1836, alr1612	COG1836	In progress in Synechocystis	
15	Pyridoxal phosphate enzyme in amino acid metabolism, most likely in the Glu-Pro area	At4g26860, At1g11930	COG0325, yggS	PROSC	In progress in <i>E. coli</i>	

Table 1. Status of most advanced fifteen families

COG3533 genes are found in all plants and occur sporadically in plant pathogens (bacteria and fungi) and in human pathogens. The corresponding proteins are similar to glycosyl hydrolase but the specific substrates are not known. The *Arabidopsis* COG3533 genes (At5g12950 and At5g12960) are expressed highly in pollen. Bacterial COG3533 genes are physically clustered with genes for hydroxyproline degradation, arabinose catabolism, and peptidases. We therefore propose that COG3533 proteins are glycosylhydrolases that cleave the hydroxyproline-linked galactosides found in plant cell wall proteins or in collagen. Such a hydrolase has been predicted to have a role in pollen growth and would allow plant pathogens to utilize plant cell wall components as carbon sources.

## 228

### Functional Annotation of Putative Enzymes in *Methanosarcina acetivorans*

Ethel Apolinario,<sup>1</sup> Libuse Brachova,<sup>3</sup> Yihong Chen,<sup>2</sup> Zvi Kelman,<sup>2</sup> Zhuo Li,<sup>2</sup> Basil J. Nikolau,<sup>3</sup> Lucas Showman,<sup>3</sup> Kevin Sowers,<sup>1</sup> and **John Orban**\* (orban@umbi.umd.edu)

<sup>1</sup>Center of Marine Biotechnology, University of Maryland Biotechnology Institute, Baltimore; <sup>2</sup>Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville; <sup>3</sup>W. M. Keck Metabolomics Research Laboratory, Iowa State University, Ames

**Project Goals: The goal of the project is to develop rapid experimental approaches for accurate annotation of putative enzymatic functions. Targets of interest range from those with tentatively assigned function to hypotheticals.**

Methane-producing organisms provide an efficient and cost-effective biofuel which is self-harvesting and can be distributed readily using infrastructure that is already in place. As with other genomes, however, accurate functional

annotation in methanogens lags significantly behind the large body of sequence data, representing a sizable gap in our understanding of biology in these organisms. We are using the methanogenic archaeon, *Methanosarcina acetivorans* (MA), as a model system for developing experimental tools for rapid and reliable annotation and validation of function. The target genes are putative enzymes in MA with detectable *in vivo* expression.

Our experimental approach utilizes a combination of methods for rapid function assignment. NMR spectroscopy is used to screen for putative substrates, products, or their structural analogs. Where possible, we have followed up on function assignments by checking to see if the MA gene can complement the corresponding *E. coli* knockout. We have used this approach to both validate and correct functional assignments in MA target genes, as will be illustrated with examples. Further, insights into the functional annotation of “hypotheticals” are being obtained by integrating mass spectrometry based metabolite profiles of gene knockouts with NMR-based approaches and these will also be discussed.

## 229

### Robust Prediction of Protein Localization Via Integration of Multiple Data Types

**Margaret Romine**<sup>1\*</sup> (Margie.romine@pnl.gov), Lee Ann McCue,<sup>1</sup> Gretta Serres,<sup>2</sup> Tatiana Karpinets,<sup>3</sup> Mustafa Syed,<sup>3</sup> Sam Purvine,<sup>1</sup> Michael Lueze,<sup>3</sup> Guruprasad Kora,<sup>3</sup> Denise Schmoyer,<sup>3</sup> Ed Uberbacher,<sup>3</sup> Jim Fredrickson,<sup>1</sup> and Mary Lipton<sup>1</sup>

<sup>1</sup>Pacific Northwest National Laboratory, Richland, Wash.;

<sup>2</sup>Marine Biological Laboratory, Woods Hole, Mass.; and

<sup>3</sup>Oak Ridge National Laboratory, Oak Ridge, Tenn.

**Project Goals: The primary research emphasis will be on associations between autotrophic and heterotrophic microorganisms with the additional objective of obtaining a predictive understanding of how interactions impart stability and resistance to stress, environmental fitness, and functional efficiency.**

Genome annotations play a central role in omics-based characterization of cellular behavior and consequently it is important that they are as accurate and functionally descriptive as possible. Currently, domain content is the primary type of functional evidence used for automated functional annotation of protein-coding genes (CDSs) deduced from genome sequences. While domain content can sometimes suggest a precise function or at least provide a general functional categorization (e.g. TonB-dependent receptor), they are more often only useful for establishing that proteins having the same domain (s) are somehow functionally related. Protein localization prediction is a form of evidence that is generally under-utilized in automated annotation pipelines but has the potential to provide very useful clues regarding CDS function. As part of our efforts to manually improve the annotation of the currently available *Shewanella* genome

sequences, we developed a strategy for more accurately predicting subcellular protein localization through integration of proteome data, the output of several different localization prediction tools, ortholog analysis, and domain analyses.

At the outset of this exercise we recognized that one of the major limitations of tools that computationally predict protein localization is that an accurate gene model is needed. Since many of the commonly used tools search for the occurrence of characteristic N-terminal targeting peptides, they will be unable to detect secretion substrates that are encoded by genes with start codon inaccuracies, gene sequencing mistakes, or genomic mutations that result in displacement or loss of sequences that encode the N-terminal targeting peptide. In order to address issues with the accuracy of the gene models we first mined available MS-MS proteome data from 12 sequenced shewanellae genomes for partial tryptic peptides that could be mapped to the mature termini of proteins deduced from the original or subsequently adjusted gene models. These analyses included searches for peptides that map to N-termini produced by cellular proteolytic processing by signal peptidase I, methionine aminopeptidase, or proline aminopeptidase. We identified such peptides for 1290 proteins (~30% of the total predicted) in the extensively studied model organism *S. oneidensis* MR-1 and between 299 and 661 proteins (~10% of the total predicted) in 11 other shewanellae for which proteome data was available but was derived from only a single sample. The positions of mobile elements (insertion elements, MITES, phage, and other integrative elements) were mapped to facilitate detection of gene fragments encoding targeting peptides that were displaced by gene interruption. This analysis resulted in an increase in pseudo-gene count from 735 to 1499. Ortholog tables comprised of proteins from all 19 strains were constructed so that we could compare, within each ortholog group, the output of several localization and domain predictors with the expectation that inconsistencies in predictions would most often arise due to errors in either the gene model or predicted ortholog grouping. Ortholog groups with inconsistencies in predicted domain content, function, or location prediction or for which members were missing in a genome were then manually evaluated for inaccuracies in gene models or ortholog grouping. These analyses lead to the addition of 769 new genes and removal of 1554 genes from the gene models of these 19 shewanellae. Taking into account only changes made to the gene models of intact genes, we adjusted the start position positions in 2466 genes thereby achieving a greater consistency in predictions of localization or domain content within each ortholog group.

Since Gram negative bacteria like *Shewanella* sp. have a complex cell envelope consisting of inner and outer membranes that are separated by a periplasmic space, they employ specialized systems to mediate translocation of proteins across one or both membranes, insertion of proteins into one membrane or the other, or to tether them to one side of a membrane. The sorting signals recognized by these systems differ from one another and thus no single algorithm is optimal for predicting the subcellular locations of all proteins. This need to apply more complex

logic for predicting protein location became evident when we discovered that representatives of all six specialized protein translocation systems (T1SS-T6SS) known to occur in gram negative bacteria were present in at least one sequenced *Shewanella*. We developed a series of rules to identify substrates of specialized secretion systems as either bioinformatics tools were not available to identify their substrates or their predictions were not particularly robust. For example, combining domain information and proteomics data for the NiFe hydrogenase orthologs allowed us to identify these proteins as substrates of the twin arginine translocation (TAT) system. Substrates of the TAT secretion system are expected to include proteins that possess metallic redox active centers and therefore all proteins having such domains, including the NiFe hydrogenases, were carefully evaluated for the presence of N-terminal targeting peptide recognized by this secretion pathway. In *Shewanella*, the NiFe hydrogenases have an unusually long targeting peptide that was validated by proteome analysis (68 amino acids) but routinely missed by both TatP and Tatfind algorithms. The identification of outer membrane proteins was also not very accurate using a single computational tool. The Bomp beta barrel prediction tool, for example, inconsistently detected outer membrane proteins within ortholog groups even after gene model adjustment. Therefore, we supplemented these analyses by searching for a C-terminal outer membrane targeting consensus motif. Since it is known that some outer membrane proteins do not encode this domain at the C-terminus (e.g. OmpA family proteins, secretins) we also used location-informative domains to assist in identification of outer membrane proteins. Other systems, such as the type II secretion system (T2SS) that translocate periplasmic proteins across the outer membrane have no universally recognized targeting motif, but are instead believed to be recognize targeting signals that are species-specific. In *Shewanella* it is known that at least three lipoproteins are substrates of this system. A comparative analysis of these lipoproteins with other proteins deduced from the genome sequence revealed a putative targeting motif similar to those described for extracellular proteins in other bacteria, providing us a means to expand the number of predicted T2SS substrates in this Genus.

We estimate that approximately 40% of the predicted proteome for each strain of *Shewanella* is translocated out of the cytoplasm. These extracytoplasmic proteins play a central role in modulating the interactions of members of this genus with their external environments and in generating the energy and accessing the nutrient necessary to support growth and metabolism. As part of PNNL's new Foundational Science Focus area on Biological Systems Interactions we intend to employ this general strategy to identify secreted proteins in new model organisms and microbial communities to facilitate future studies directed at developing a broader understanding of microbial interactions.

submitted post-press

## Genome-Scale Phylogenetic Function Annotation of Large and Diverse Protein Families

Barbara E. Engelhardt,<sup>1,4</sup> Michael I. Jordan,<sup>2</sup> Susanna Repo,<sup>3</sup> and Steven E. Brenner<sup>3</sup> (brenner@compbio.berkeley.edu)

<sup>1</sup>EECS Dept., <sup>2</sup>Dept. of Statistics, and <sup>3</sup>Plant and Microbial Biology Dept., University of California, Berkeley; and <sup>4</sup>Computer Science Dept., University of Chicago, Ill.

**Project Goals: The goal of the project is to enhance the algorithms and statistical models of SIFTER, our protein function prediction method. We will also extend SIFTER's applicability by including additional sources of function evidence. With these improvements, SIFTER will become applicable to a broader range of protein families, including large, and functionally diverse families, and to work on genome-wide scale. In addition, we will adapt SIFTER on metagenomic data.**

It is now easier to discover thousands of protein sequences in a new microbial genome than it is to biochemically characterize the specific activity of a single protein of unknown function. Through metagenomic analysis, next-generation sequencing heralds unprecedented opportunities for understanding the environmental microbiota. A single experiment alone, the Global Ocean Sampling study, more than doubled the number of known protein sequence entries. However, despite this large body of new sequence information, functional annotation remains a major challenge. Molecular functions of proteins in the novel genomes continue to be discovered, in large part by homology to those experimentally characterized in model organisms.

Typically, protein function annotation involves finding homologs of a protein sequence, followed by database queries and computational techniques to predict function from the annotated homologs. These methods rely on the principle that proteins from a common ancestor may share a similar function. However, most protein families have sets of proteins with different functions and therefore traditional bioinformatics approaches are unable to reliably assign the appropriate function to unannotated proteins. Currently, protein function databases have a large proportion of erroneously annotated proteins, where the incorrect annotations were either derived using an imprecise computational technique or inferred using another incorrect annotation<sup>1-4</sup>.

We have proposed integrating available functional data using the evolutionary relationships of a protein family, and we implemented this method in the program SIFTER (Statistical Inference of Function Through Evolutionary Relationships). The SIFTER methodology uses a statistical graphical model to compute the probabilities of molecular functions for unannotated proteins. Currently, SIFTER takes as input a reconciled phylogeny and a set of annota-

tions for some of the proteins in the protein family. We incorporate known information about function by computing the probability of each of the candidate functions for the proteins in the tree with available functional evidence from the GOA database. The candidate molecular functions are represented as a boolean vector, where initially the probability associated with each candidate function is a function of the set of annotations for that protein and their corresponding evidence types (e.g., experimental, electronic). From this reconciled phylogeny with sparse observations, SIFTER computes the posterior probability of each molecular function for all proteins in the family using a simple statistical model of protein function evolution.

We tested the performance of SIFTER on three different protein families: AMP/adenosine deaminases, sulfotransferases and Nudix hydrolases with cross-validation experiments. SIFTER's performance was compared with three other function prediction algorithms: BLAST, GOtcha and Orthostrapper, and SIFTER was shown to outperform the other methods. In addition, on a genome-wide scale we used SIFTER to annotate the experimentally characterized proteins from *Schizosaccharomyces pombe*, based on the annotations from 26 other fungal genomes. The newest version of SIFTER implements a faster method for calculating the posterior probabilities, and this improvement, together with a more general evolutionary model make SIFTER applicable on large and functionally diverse protein families and on genome-scale function annotation.

The development of SIFTER is an ongoing project and a new version of the program is now available (manuscript under review). We are currently testing SIFTER for metagenomic sequences with the acid mine drainage datasets from Jill Banfield. In the near future, we are planning to expand our analysis to other metagenomic datasets, such as the termite gut datasets from the JGI. We also use SIFTER to annotate enzymes from chlorite dismutase and perchlorate reductase families, in order to identify species that are capable of perchlorate reduction. Furthermore, we are validating SIFTER predictions experimentally using the large and extremely diverse Nudix family of hydrolases as a test bed.

This project has been funded with DOE grant number BER KP 110201.

## References

1. Brenner SE 1999 *Trends Genet.* **15** 132-3
2. Galperin MY and Koonin EV 1998 *In Silico Biol.* **1** 55-67
3. Jones CE, Brown AL and Baumann U 2007 *BMC Bioinformatics.* **8** 170
4. Schnoes AM, Brown SD, Dodevski I and Babbit PC 2009 *PLoS Comput Biol.* **5** e1000605

## Computing for Systems Biology

# 230

## Standards in Genomic Sciences: Launch of a Standards Compliant Open-Access Journal for the 'Omics Community

**G. M. Garrity**<sup>1,5\*</sup> (garrity@msu.edu), N. Kyrpides,<sup>2,5</sup> D. Field,<sup>3,5</sup> P. Sterk,<sup>3,5</sup> H.-P. Klenk,<sup>4,5</sup> and the Editorial and Advisory Boards of Standards in Genomic Sciences

<sup>1</sup>Michigan State University, East Lansing; <sup>2</sup>DOE Joint Genome Institute, Walnut Creek, Calif.; <sup>3</sup>NERC Center for Ecology and Hydrology, Oxford, United Kingdom; <sup>4</sup>DSMZ – German Collection of Microorganisms and Cell Cultures GmbH, Braunschweig, Germany; <sup>5</sup>Genomic Standards Consortium, Seattle, Wash.

**Project Goals: The goal of DOE funding was to underwrite a pre-launch meeting of the Editorial and Advisory Boards of Standards in Genomic Sciences. The meeting was held at Michigan State University on March 12–14, 2009.**

*Standards in Genomic Sciences* (SIGS) is an open-access e-journal that was created to promote the data standardization efforts of the Genomic Standards Consortium (GSC). The GSC was founded in 2005 by an international community of like-minded scientists to work toward improving the descriptions of our rapidly growing collection of genomes and metagenomes [1,2]. In the absence of metadata standards, the difficulty of exchanging and integrating genomic data into analytical models and public knowledgebases increases while the overall value of each subsequent sequence diminishes. This is problematic because the ease and cost of producing sequence data have dropped sharply while the cost of annotation and documentation has increased.

Membership in the GSC consists of biologists, bioinformaticians, and computer scientists, with representatives from the National Center for Biological Information (NCBI), European Molecular Biology Laboratory (EMBL), National Institute of Genetics Japan (NIG), J. Craig Venter Institute (JCVI), DOE Joint Genome Institute (JGI), European Bioinformatics Institute (EBI), Sanger Institute, and a number of other international research organizations involved in cross-cutting research. As a first step toward accomplishing organizational goals, the GSC published the "Minimum Information about a Genome Sequence" (MIGS) specification, which describes the core information that should be reported with each new genome or metagenome [3]. The GSC has led the development of a richer set of descriptors within GCDML (Genomic Contextual Markup Language), an XML variant for mark-up and transport of genomic and metagenomic data and the Genomic Rosetta Stone, a proposed resolver for mapping genome identifiers across databases [4,5]. The GSC also participates in initiatives on data

standardization, including Biosharing and the sequencing finishing standards recently described by Chain et al. [6,7]

The rationale for SIGS is to provide a venue for publication of highly structured, standardized publications of genome and metagenome sequences in accordance with MIGS and to report on other efforts that promote data standardization and data sharing [8]. Whereas peer-reviewed publications of genomes were once commonplace in a number of journals, the current trend is for many general and discipline-specific publications to eschew such papers, leading to a loss of contextual information that is critical for analyzing and interpreting genome sequence data [9]. SIGS aims to counter this trend and to provide concise, standardized descriptions of the sequencing and annotation methods along with biological information about the source organism, which we refer to as short genome reports. To that end, and with the generous support from the Michigan State University Foundation to fund the editorial office and the U.S. Department of Energy Office of Science, Biological and Environmental Research Program to convene the first meeting of the editorial and advisory boards, an open-access publication was launched to help meet those needs.

Publication of SIGS began in July 2009. At the end of October, the journal had published 28 short genome reports on bacterial and archaeal species, many of which were derived from the *Genomic Encyclopedia of Bacteria and Archaea* collaboration between the DOE Joint Genome Institute (JGI) and the German Collection of Microorganisms and Cell Cultures (DSMZ). We anticipate publishing at least another 14 short genome reports before the close of 2009, along with approximately five to six additional papers, bringing the total to approximately 60 published articles in the first volume.

Growth of the journal, to date, has been largely organic, through the journal website, search engines and forward linking of SIGS DOIs on the websites of other journals that have been cited in SIGS articles. Published articles have been downloaded > 4,750 times. Readership of SIGS is worldwide, with visitors to the site coming from over 60 countries during November, with the preponderance of visitors coming from North America and western Europe. Approximately half the daily visitors are new to the site, and the bulk of that traffic appears to be directed to the site either by search engines or by direct linking from other sites. We anticipate that traffic will continue to grow as additional content is published and SIGS becomes accepted in the major literature indices (PubMed Central, PubMed, ISI Web of Science). Our goal is to engage with the GTL community to solicit feedback and discuss additional unmet publishing needs.

## References

1. Field D., Hughes J. Cataloguing our Current Genome Collection. *Microbiology* 2005; 151: 1016-1019. PubMed doi:10.1099/mic.0.27914-0.
2. Field D., Garrity G., Morrison N., Sterk P., Selengut J., Thomson N., Tatusova T. Meeting Report: eGenomics: Cataloguing Our Complete Genome Collection I. *Comp Funct Genomics* 2006; 6: 357-362. doi:10.1002/cfg.493.

3. Field D., Garrity G., Gray T., Morrison N., Selengut J., Sterk P., Tatusova T., Thomson N., Allen M.J., Angiuoli S.V., et al. The Minimum Information about a Genome Sequence (MIGS) Specification. *Nat Biotechnol* 2008; 26: 541-547. PubMed doi:10.1038/nbt1360.
4. Kottmann R., Gray T., Murphy S., Kagan L., Kravitz S., Lombardot T., Field D., Glockner F.O. A Standard MIGS/MIMS Compliant XML Schema: Toward the Development of the Genomic Contextual Data Markup Language (GCDML). *OMICS* 2008; 12: 115-121. PubMed doi:10.1089/omi.2008.0A10.
5. Van Brabant B., Gray T., Verslyppe B., Kyrpides N., Dietrich K., Glockner F. O., Cole J., Farris R., Schriml L. M., De Vos P., et al. Laying the Foundation for a Genomic Rosetta Stone: Creating Information Hubs Through the Use of Consensus Identifiers. *OMICS* 2008; 12: 123-127. PubMed doi:10.1089/omi.2008.0020.
6. Chain P. S., Grafham D.V., Fulton R. S., Fitzgerald M. G., Hostetler J., Muzny D., Ali J., Birren B., Bruce D. C., Buhay C., et al. Genomics. Genome Project Standards in a New Era of Sequencing. *Science* 2009; 326: 236-237. PubMed doi:10.1126/science.1180614.
7. Field D., Sansone S. A., Collis A., Booth T., Dukes P., Gregurick S. K., Kennedy K., Kolar P., Kolker E., Maxon M., et al. Megascience. 'Omics Data Sharing. *Science* 2009; 326: 234-236. PubMed doi:10.1126/science.1180598.
8. Garrity, G. M., Field D., Kyrpides N., Hirschman L., Sansone S. A., Angiuoli S., Cole J. R., Glockner F.O., Kolker E., Kowalchuk G., et al. Toward a Standards-Compliant Genomic and Metagenomic Publication Record. *OMICS* 2008; 12: 157-160. PubMed doi:10.1089/omi.2008.A2B2.
9. Liolios K., Mavromatis K., Tavernarakis N., Kyrpides N. The Genomes On Line Database (GOLD) in 2007: Status of Genomic and Metagenomic Projects and Their Associated Metadata. *Nucleic Acids Res* 2008; 36: D475-D479. PubMed doi:10.1093/nar/gkm884.

# 231

## NamesforLife Semantic Resolution Services for the Life Sciences

**Charles T. Parker**,<sup>1</sup> Dorothea Rohlfis,<sup>1</sup> Sarah Wigley,<sup>1</sup> Nicole Osier,<sup>1</sup> Catherine Lyons,<sup>1</sup> and George M. Garrity<sup>1,2,\*</sup> (garrity@namesforlife.com)

<sup>1</sup>NamesforLife, LLC, East Lansing, Michigan and  
<sup>2</sup>Michigan State University, East Lansing

**Project Goals: The overall objective of the Phase II study is to develop and deploy a set of convenient, easy to use semantic services that provide end-users with on-demand access to key information. This ensures that they can accurately interpret the meaning of any bacterial or archeal name when encountered in digital content.**

Within the Genomes-to-Life Roadmap, the DOE recognizes 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 interpre-

tation of older data and published results decreases because both become overloaded with synonymous (multiple terms for a single concept) and polysemous terms (single terms with multiple meanings). Ambiguity arising from rapidly evolving terminology is a common and chronic problem in science and technology. N4L services are being developed to address this problem. The core of N4L consists of a data model, an XML schema, and an expertly managed vocabulary that is interlinked with Digital Object Identifiers (DOIs) to form a transparent semantic resolution service that disambiguates terminologies, makes them actionable, and provides direct links back to key literature and data resources.

**Objectives** - The overall objective of the current Phase II study is to deploy a set of convenient, easy to use semantic services that allow end-users to accurately interpret the meaning of a biological name or other dynamic term encountered in digital content, on demand and without having to query external resources or to leave the material they are reading or searching. The service can be used by database owners, publishers, or other information providers, to semantically enable their offerings; making them readily discoverable by their clients, even when the definition of a name or term has changed.

**Curatorial Efforts** - We significantly extended the scope of our data curation and built a framework for distributing and enhancing N4L information services to different categories of users. The target vocabulary consists of the validly published names of Bacteria and Archaea, which provides a rich and complex set of interrelated terms and interlinked resources that have a high value to the GTL community. At the end of 4Q 2009, there were 13043 validly published names (of which 3022 are synonyms of varying complexity) corresponding to 12630 taxonomic concepts and 8976 biological entities. Currently, new validly published names appear in the literature at a rate of 3.9 names/day. This number is however significantly lower than the number of names that have no standing in the literature (14.9 names/day) that appear in INSDC records and the GenBank taxonomy. Trivial names appearing on INSDC records add further confusion to the process and occur at a rate more than five-fold higher than validly published names.

The NamesforLife data have undergone further refinement to improve their accuracy. All names, taxon and exemplar records have been asserted by literature, corresponding to 9474 references, including 277 involving judicial opinions that affect the legitimacy and valid publication of 433 names. This has significance to the GTL program as some genomes are currently posted under rejected names (e.g. *Sinorhizobium medicae*). We have also verified all of the strains, culture collection deposits, and 16S rRNA sequences used in taxonomic assertions based on a review of the published record. This addresses a growing problem that has arisen from more than a decade of automated data harvesting, coupled with transitive data closure, leading to numerous systematic errors that are being routinely re-propagated.

**N4L BrowserTool** - The N4L BrowserTool provides a means of wide-spread distribution of our semantic

resolution services to end-users of scientific and technical literature, published in digital form and distributed via the web. The tool is currently distributed as a Firefox extension and provides on-the-fly client-side mark-up of bacterial and archaeal names with links to NamesforLife information objects that can be actuated on demand. Alpha testing of the BrowserTool ran from May - December 2009. Large-scale beta-testing is scheduled for January-February 2010 with a product release in March 2010.

**N4L Autotagger** - The N4L Autotagger provides publishers and other content providers with a method for enabling and enhancing content during composition. This results in articles that contain persistent links to N4L information objects and allow readers to view such content in any browser. Collaborative work is underway with the Society for General Microbiology to enhance and enable content appearing in the International Journal of Systematic and Evolutionary Microbiology.

**N4L Contextual Index** - The BrowserTool and AutoTagger are designed to recognize bona-fide nomenclatural events in pre-composition XML and HTML, thus allowing for high-fidelity harvesting of new/modified names and associated references from the taxonomic literature automatically. These tools can also capture the metadata for each source, thus allowing us to track all such events. This information is being used to create a contextual index that enhances the value of N4L tools as each successive use can be placed into a variety of larger contexts and used for a variety of purposes, ranging from resource planning to plotting research trends at both a fine-grained (taxon specific) and global level. In addition to the scientific literature, we are building the necessary infrastructure to permit the use of our tools to uncover prior art in areas of interest to the DOE (e.g. bio-energy/biobased feedstocks/genomics) in the U.S. and EPO patent literature.

This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy under Phase II STTR Award DE-FG02-07ER86321 A001

## 232

### Numerical Optimization Algorithms and Software for Systems Biology: An Integrated Model of Macromolecular Synthesis and Metabolism of *Escherichia coli*

Ines Thiele<sup>1\*</sup> (ithiele@hi.is), R.M.T. Fleming,<sup>1</sup> A. Bordbar,<sup>2</sup> R. Que,<sup>2</sup> and B.O. Palsson<sup>2</sup>

<sup>1</sup>Center for Systems Biology, University of Iceland, Reykjavik, Iceland and <sup>2</sup>Bioengineering Dept., University of California, San Diego, La Jolla

**Project Goals:** This project aims to reconstruct a genome-scale model of metabolism and macromolecular synthesis and to develop algorithms capable of solving the resulting large, stiff and ill-scaled matrices. This project combines

**state of the art reconstruction and constraint-based modeling analysis tools with high-end linear optimization solvers and convex flux balance analysis. The incorporation of thermodynamic information in addition to environmental constraints will allow an accurate assessment of feasible steady states. While we will prototype the reconstruction and algorithm developments with *Escherichia coli*, we will employ the resulting networks to determine thermodynamically favorable pathways for hydrogen production by *Thermotoga maritima*.**

Systems biology is a rapidly growing discipline. It is widely believed to have a broad transformative potential on both basic and applied studies in the life sciences. In particular, biochemical network reconstructions are playing a key role as they provide a framework for investigation of the mechanisms underlying the genotype-phenotype relationship. The constraint-based reconstruction and analysis approach was applied to reconstruct the transcriptional and translational (tr/tr) machinery of *Escherichia coli*. This reconstruction, denoted 'Expression-matrix' (E-matrix), represents stoichiometrically all known proteins and RNA species involved in the macromolecular synthesis machinery. It accounts for all biochemical transformations to produce active, functional proteins, tRNAs, and rRNAs known to be involved in *E. coli*'s tr/tr machinery. An initial study investigated basic properties of the E-matrix, including its capability to produce ribosomes, which was found to be in good agreement with experimental data from literature. Furthermore, quantitative gene expression data could be integrated with, and analyzed in the context of, the resulting constraint-based model. Adding mathematically derived constraints to couple certain reactions in the model allowed the quantitative representation of the size of steady state protein and RNA pools.

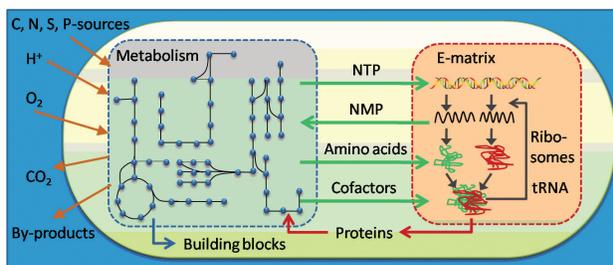


Figure 1: Functional synergy between the metabolic network and the macro-molecular synthesis network in *E. coli*.

The E-matrix was integrated with the genome-scale *E. coli* metabolic model and extended the transcriptional and translational reactions to encompass genes encoding all the respective metabolic enzymes. The resulting Metabolite-Expression-matrix (ME-matrix), exceeds the predictive capacity of the metabolic model and it can, for example, be used to predict the biomass yield since it represents the production of almost 2,000 proteins. *E. coli*'s ME-matrix is the first of its kind and represents a milestone in systems biology as it demonstrates how to quantitatively integrate 'omics'-datasets into a network context, and thus, to study the mechanistic principles underlying the genotype-phenotype

relationship. We will show some possible applications which include protein engineering, interpretation of adaptive evolution, and minimal genome design. An integration of the ME-matrix with remaining cellular processes, such as regulation, signaling, and replication, will be a next step to complete the first whole-cell model.

Building on this reconstruction effort we now started to construct the ME-matrix for *Thermotoga maritima* based on published data. Furthermore, significant advances have been made in incorporating thermodynamic constraints with metabolic networks, as shown in the accompanying poster "Numerical Optimization Algorithms and Software for Systems Biology". This work sets the stage for the goal of thermodynamically favorable pathways for hydrogen production by *Thermotoga maritima*.

## 233

### The Ribosomal Database Project: Tools and Sequences for rRNA Analysis

J.R. Cole\* (colej@msu.edu), Q. Wang, B. Chai, J. Fish, E. Cardenas, R.J. Farris, D.M. McGarrell, G.M. Garrity, and J.M. Tiedje

Michigan State University, East Lansing

**Project Goals: The Ribosomal Database Project (RDP; <http://rdp.cme.msu.edu>) offers aligned and annotated rRNA sequence data and analysis services to the research community. These services help researchers with the discovery and characterization of microbes important to bioenergy production, biogeochemical cycles, including those affecting carbon and climate, and bioremediation.**

Updated monthly, the RDP maintained 1,281,097 aligned and annotated quality-controlled rRNA sequences as of December 2009 (Release 10.17). The *myRDP* features have grown to support a total of over 2,500 active researchers using their *myRDP* accounts to analyze over 4,000,000 pre-publication sequences in 38,708 sequence groups, while the RDP Pyrosequencing Pipeline has been used by over 700 researchers to analyze next-generation sequences.

#### New NCBI/EMBL Short Read Archive Submission Tool:

Because it is very difficult for researchers to submit their next-generation rRNA sequence data to the three INSDC databases (GenBank, EMBL, and DDBJ), RDP developed a combination of web and downloadable programs, the *myRDP* SRA PrepKit, to allow users to prepare and edit their submissions. This package provides an effective solution to the difficult and confusing process involved in preparing metadata documents that are required for submission to the GenBank Short Read Archive (SRA) or EMBL European Read Archive (ERA), the two databases for reads generated from ultra-high-throughput sequencing technologies. It can be applied on sequence data generated from 454 (GS 20, FLX and Titanium), Illumina/Solexa, ABI SOLiD and Helicos platforms. It transforms the

preparation of six separate XML document types required for each submission into a clear flow of tasks implemented in easy-to-understand forms for collecting metadata about the study, samples, experiments, analyses, and sequencing runs. A set of suggested attributes in the data forms assist researchers in providing metadata conforming to the MIMS Minimal Information about a Metagenome Sequence specification, and the upcoming MEINS Minimal Information about an Environmental Sequence specifications (Field et al., 2008, *Nat. Biotechnol.* 26:541; [http://gensc.org/gc\\_wiki/index.php/MIENS](http://gensc.org/gc_wiki/index.php/MIENS)). The user can save unfinished work for later sessions and copy individual components to new submissions to avoid repetitive entry of shared data. In addition, a provided Java Web Start program creates a Fastq file from sequence reads. The *myRDP* Submission Web Start program makes it easy to perform the tasks needed to finalize your submission. A help page outlining the workflow is also provided. (The USDA provided additional funding for the *myRDP* SRA PrepKit.)

**RDP Pyrosequencing Pipeline:** This toolkit has been used by 777 researchers (unique e-mail addresses) to analyze their next-generation sequence data. This pipeline offers a collection of tools that automate the data processing and simplify the computationally intensive analysis of large sequencing libraries. A number of new functions have been developed for the pipeline, including a new distance matrix tool that generates distance matrices in two popular formats used by third-party tools such as Mothur (Schloss, 2009, *Appl. Environ. Microbiol.* 75:7537). All the tools now accept compressed files to reduce the upload time of large amounts of sequence data. The Initial Process, Aligner, and Clustering tools have been enhanced to return graphical summary files that provide a visual representation of sequence quality and diversity. (The USDA and NIEHS provided additional support for the RDP Pyrosequencing Pipeline.)

Other RDP tools have been used, on average, in **18,633 analysis sessions per month** by an average of **5,634 researchers** (unique IPs). These include the **RDP Classifier**, which is also available as an open-source package through SourceForge and has been **downloaded 729 times**, the online Infernal secondary-structure based aligner (Nawrocki, 2009, *Bioinformatics* 25:1335) trained by RDP on representative bacterial and archaeal alignments, the **RDP Sequence Match** program for finding nearest neighbors, the **RDP Library Compare** program for determining differentially represented taxa between two environmental libraries, the **RDP Probe Match** program for determining taxonomic coverage of primers and probes, the **RDP Tree Builder** for rapid phylogenetic tree construction, and the **RDP Hierarchy Browsers** that provide entry to the RDP sequences in taxonomic order, by publication, or by completed genome (many genomes contain multiple rRNA operons). A **new RDP Multi-Classifer** is being provided as a command-line tool to accommodate the growing need for taxonomy-based analyses of large numbers of sequences in multiple samples. This tool combines the functions of both RDP Classifier and Library Compare, and thus provides a convenient solution for researchers to use as standalone tools or to be integrated into their own analysis workflow.

**RDP Web Services** have been expanded to provide interfaces for the RDP Classifier, Sequence Match, Probe Match and *myRDP* tools. There are, on average, **198,632 SOAP requests** received per month. Usage examples are provided in Java and Ruby. Researchers can incorporate these web services into their own analysis pipelines to make use of these popular RDP tools.

This research is supported by the Office of Science (BER), U.S. Department of Energy under Grant No. DE-FG02-99ER62848.

## References

1. Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R.J. Farris, A.S. Kulam-Syed-Mohideen, D.M. McGarrell, T. Marsh, G.M. Garrity, and J.M. Tiedje. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* 37 (Database issue): D141-D145; doi:10.1093/nar/gkn879.
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.* 73:5261-7; doi:10.1128/AEM.00062-07.

# 234

## Identifying Proteins from Microbial Communities

**William R. Cannon**<sup>1\*</sup> (William.Cannon@pnl.gov), Mitchell Rawlins,<sup>1</sup> Gaurav Kulkarni,<sup>2</sup> Andy Wu,<sup>2</sup> Ananth Kalyanaraman,<sup>2</sup> Douglas Baxter,<sup>1</sup> Mary Lipton,<sup>1</sup> and Steven Callister<sup>1</sup>

<sup>1</sup>Pacific Northwest National Laboratory, Richland, Wash. and <sup>2</sup>Washington State University, Pullman, Wash.

**Project Goals: See below.**

The lack of reliable genome sequences currently limits the effectiveness of proteomics studies of microbial communities because of the difficulty in identifying peptides. Characterizing the proteomics of microbial communities requires (1) the computational interpretation and integration of high-throughput experimental data, (2) the leveraging of existing sources of knowledge from multiple domains, and (3) searching for solutions that meet criteria on multiple levels in a large search space. Our goal is to develop novel methods needed to describe the proteins and metagenomic functional processes occurring within unsequenced microbial communities being investigated as part of DOE's missions in carbon sequestration, bioremediation and bioenergy research.

## 235

Student Presentation

### Identifying the Mediators of Environmental Changes Through Integration of Steady State and Time-Course Gene Expression Profiles in *Shewanella oneidensis* MR-1

Qasim K. Beg,<sup>1</sup> Mattia Zampieri,<sup>2,5</sup> Sara Baldwin<sup>2\*</sup> (baldwin2@bu.edu), Niels Klitgord,<sup>2</sup> Margrethe H. Serres,<sup>4</sup> Claudio Altafini,<sup>5</sup> and Daniel Segre<sup>1,2,3</sup>

<sup>1</sup>Dept. of Biomedical Engineering, <sup>2</sup>Bioinformatics Program, and <sup>3</sup>Dept. of Biology, Boston University, Boston, Mass.; <sup>4</sup>Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, Mass.; and <sup>5</sup>International School for Advanced Studies, Trieste, Italy

**Project Goals:** In this study, we combine mRNA microarray and metabolite measurements with statistical inference and dynamic flux balance analysis to study the transcriptional response of *S. oneidensis* as it passes through exponential, stationary, and transition phases. By measuring time-dependent mRNA expression levels during batch growth of *S. oneidensis* MR-1 under two radically different nutrient compositions, we obtain detailed snapshots of the regulatory strategies used by this bacterium to cope with gradually decreasing nutrient availability.

The dynamics of transcriptional regulation in microbial growth is an environment-dependent process. This dynamics is strongly controlled by two main factors: the wiring of the underlying regulatory network, and the time-dependent array of environmental stimuli. Understanding the interplay between these two factors is a fundamental challenge in systems biology, particularly relevant for the study of microbial systems, often adapted to rapidly changing environments. Certain genes may be activated as a response to the lack of a specific nutrient, and therefore display a strong dependence on environmental conditions; others may be more generally associated with growth rate, or growth phase requirements, and could therefore show similar behavior across different media. We address these questions in the environmental microbe *Shewanella oneidensis* MR-1, whose versatile respiratory functions make it a key player in environmental and bioenergy applications.

In this study, we combine mRNA microarray and metabolite measurements with statistical inference and dynamic flux balance analysis to study the transcriptional response of *S. oneidensis* as it passes through exponential, stationary, and transition phases. By measuring time-dependent mRNA expression levels during batch growth of *S. oneidensis* MR-1 under two radically different nutrient compositions (minimal lactate medium and LB medium), we obtain detailed snapshots of the regulatory strategies used by this bacterium to cope with gradually decreasing nutrient availability. In addition to traditional clustering, which provides a first indication of major regulatory trends and transcription factors activities, we implement a new approach for Dynamic Detection of Transcriptional Triggers (D2T2). This new

method allows us to infer a putative topology of transcriptional dependencies, with special emphasis on the nodes at which external stimuli are expected to affect the internal dynamics. In parallel, we address the question of how to compare transcriptional profiles across different time-course experiments. Our growth derivative mapping (GDM) method makes it possible to relate with each other points that correspond to the same relative growth rate in different media. This mapping allows us to discriminate between genes that display an environment-independent behavior, and genes whose transcription seems to be tuned by specific environmental factors.

Several observed transcript time-courses raise interesting biological questions. For example, we observe a coupling between nitrogen-related genes and the glycogen biosynthesis/degradation pathway. To help rationalize the observed patterns, we measure extracellular metabolites and show how transcription and metabolism can be interpreted in the context of a dynamic flux balance analysis model.

## 236

### Computational Design of Microbial Cross-Feeding Induced by Synthetic Growth Media

Niels Klitgord<sup>1\*</sup> (niels@bu.edu) and Daniel Segre<sup>1,2</sup> (dsegre@bu.edu)

<sup>1</sup>Program in Bioinformatics and <sup>2</sup>Dept. of Biology and Dept. of Biomedical Engineering, Boston University, Boston, Mass.

<http://prelude.bu.edu>

**Project Goals:** We seek to develop algorithm for engineering novel microbe-microbe interactions. Our method, based on stoichiometric genome-scale models of metabolism, is aimed at identifying environment that induce cross-feeding interactions. We envisage that such a “synthetic ecology” approach will be relevant for environmental and bioenergy applications.

Microbial ecosystems are ubiquitous on our planet, and play a major role in the global balance of the biosphere, as well as in the ongoing efforts for establishing renewable bioenergy sources. Since most microbe-microbe and microbe-environment interactions are likely mediated by metabolic intermediates, understanding the flow of metabolism between microbes constitutes a fundamental unsolved challenge. Here, towards addressing this challenge, we show how stoichiometric genome-scale models of metabolism can be extended to the ecosystem level, helping identify, understand and engineer interactions between pairs of microbial species. Specifically, we propose a novel suite of algorithms that can identify artificial environments predicted to induce mutualistic interactions between two given microbial species, by efficiently searching for growth media that sustain growth of two species only when simultaneously present. Our strategy is based on two major steps: *First*, we implement a procedure for automatically joining together the

stoichiometric models for two species, embedding them into a common environment. *Second*, we search the space of possible nutrient combinations for media that could not sustain growth of each species alone, but allow growth of both species simultaneously.

We validated our approach using three organism pairs of increasing complexity. The first is a simple toy model, in which one can arbitrarily pre-define expected mutualistic interactions. The second is a special case of the naturally occurring interactions between methanogenic archaea and hydrogen-producing microorganisms, which was recently analyzed in detail using flux balance models. The third is an experimentally engineered synthetic biological system of two yeast strains that can grow only in the presence of each other, because each of them is unable to synthesize a specific essential metabolite. In addition to recapitulating these known interactions, we will use our approach to generate new experimentally testable predictions of environments that induce interactions between pairs of environmentally relevant microorganisms, including *Shewanella oneidensis*. Selected predictions will be tested experimentally. We envisage that these algorithms will make it possible to engineer novel metabolism-based interactions between pairs of microbial species, helping develop a new computationally-driven synthetic ecology discipline

## 237

### Multi-Scale Spatially Distributed Simulations of Microbial Ecosystems

William J. Riehl<sup>\*1</sup> (briehl@bu.edu), Niels Klitgord,<sup>1</sup> Christopher J. Marx,<sup>2</sup> Nathaniel C. Cady,<sup>3</sup> and Daniel Segre<sup>1,4</sup> (dsegre@bu.edu)

<sup>1</sup>Graduate Program in Bioinformatics, Boston University, Boston, Mass.; <sup>2</sup>Dept. of Organismic and Evolutionary Biology, Harvard University, Cambridge, Mass.; <sup>3</sup>College of Nanoscale Science and Engineering, University at Albany, N.Y.; and <sup>4</sup>Depts. of Biology and Biomedical Engineering, Boston University, Boston, Mass.

<http://prelude.bu.edu>

**Project Goals:** The goals of this project are: (1) to extend current genome-scale models to include spatio-temporal dynamics, (2) to allow more realistic simulations of microbial growth for individual species and ecosystems; and (3) to enable open source development of models for the study of renewable bioenergy sources, bioremediation challenges and ecosystem balance.

Genome-scale models of microbial metabolism represent the most advanced synthesis of genomic information, biochemical knowledge, and computational efficiency relevant for developing a predictive, quantitative understanding of microbial ecosystems. These models are becoming increasingly relevant for use in a number of endeavors, such as bioenergy production, bioremediation, and carbon and nitrogen cycling in the biosphere. As automated annotation pipelines,

network gap-filling algorithms, and high throughput experimental methods improve, we will gradually approach the capacity to model virtually any sequenced microbe using this approach. Yet, some of the most fundamental properties of natural microbial ecosystems crucially depend on aspects that are well beyond the stoichiometries of individual biochemical species. These include contact- or metabolite-mediated interactions between different microbes, dynamical changes of the environment, spatial structure of the underlying geography and evolutionary competition between distinct subpopulations.

We present the early stage development of a broadly applicable and user-friendly platform for modeling these interactions by performing spatially distributed time-dependent flux balance based simulations of microbial ecosystems. We use a modified version of dynamic flux balance analysis (dFBA) to implement the dynamics of the system. By taking advantage of the computational efficiency involved in flux balance model calculations, we implement a spatially structured lattice of interacting metabolic subsystems. These subsystems represent a level of detail that is intermediate between a fine-grained single-cell modeling approach, and a broad global population modeling approach, and performs akin to a cellular automaton.

This platform has been developed with the capacity to bridge multiple spatial and temporal scales, making it possible to observe long term dynamics of microbial populations growing in a given environmental setting, based on constant updates of local nutrient availabilities and exchanges, and ultimately determined by the activity of individual metabolic reactions present in each microbial species. Thus, it can be used as a platform for modeling the spatial and temporal growth of a single bacterial species in a Petri dish, biofilm formation on complex substrate morphologies, seasonality of microbial communities in a specific geographical setting, or the growth and diffusion of a microbe that has been genetically engineered toward bioremediation in a contaminated body of water.

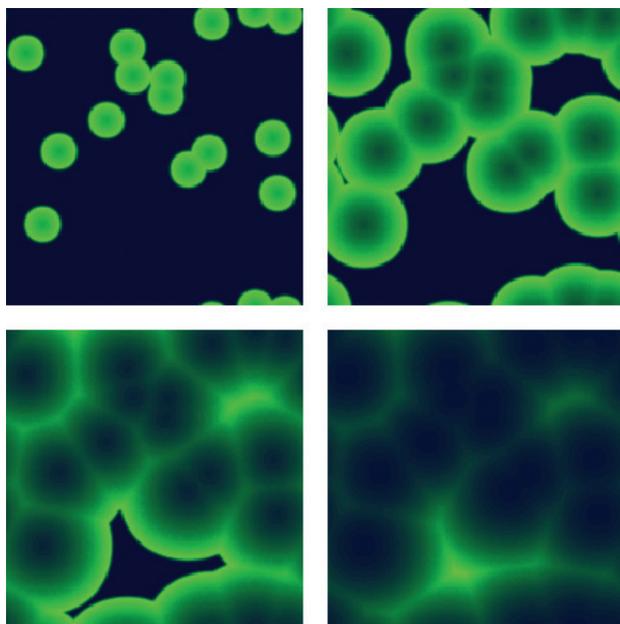


Figure 1. A sample run of the modeling platform, showing the effects of a model of *E. coli* colonies growing and merging together on a 2D surface with limited nutrient.

We present a prototype of our platform, which uses the open-source GNU Linear Programming Kit (GLPK) for performing the dFBA calculations, and a Java-based language (Processing) for coordinating the simulations and rapid visualization. We have applied this prototype to the analysis of several different examples, including the growth of a single species in a 2-Dimensional environment (see Figure 1) and syntrophic growth of microbial species. In future work, computer simulations will be integrated with experiments, allowing us to (i) calibrate the simulation parameters towards faithful representation of microbial growth patterns, and (ii) perform pilot studies on microbial ecosystem dynamics.

## 238

### Ground and Transition State Binding Calculations to Improve Cytochrome P450<sub>BM3</sub> Reactivity and Specificity

Costas D. Maranas,<sup>1</sup> George A. Khoury<sup>1\*</sup> (khoury@psu.edu), Michael J. Janik,<sup>1</sup> Patrick C. Cirino,<sup>1</sup> and Ping Lin<sup>2</sup>

<sup>1</sup>Dept. of Chemical Engineering and <sup>2</sup>Materials Simulation Center, Pennsylvania State University, University Park

<http://maranas.che.psu.edu>

**Project Goals:** The main goal of this work is to assess the impact of molecular interactions between an enzyme and its substrate at the ground and transition states on reactivity and substrate specificity. The identified trends

are currently used to inform a method for systematically re-designing Cytochrome P450<sub>BM3</sub> to hydroxylate ethane.

In this work, we introduce the combined use of ground state and transition state calculations to understand how specific mutations present in engineered variants of cytochrome P450<sub>BM3</sub> confer improved reactivity. The cytochrome P450<sub>BM3</sub> monooxygenase has been the target of extensive directed evolution by other groups. The fatty acid hydroxylase is functionally expressed at high levels in *E. coli* and has been engineered to convert small alkanes to their corresponding alcohols, with an emphasis in biofuel production. We first identified and calculated the ground and transition state structures for the rate-limiting step using quantum mechanical methods. Next, we computationally assessed the effects of 14 different experimentally isolated mutations in P450 mutant 535-h (3 mutations lie in the active site) on interactions with the ground and transition state structures with a newly developed computational saturation mutagenesis procedure. The general trend found was that some mutations are important for improving substrate binding, while other mutations in different positions are important for improving transition state stabilization. We find that calculations at both ground and transition state appear to be important for rational enzymatic design. In the design phase, we systematically chose design positions based on sequence, structure, and energetic factors, and customized the Iterative Protein Redesign and Optimization (IPRO) framework to identify the energetically optimal mutations with the ground and transition states. We report on the general trends from the optimal designs predicted by IPRO.

## 239

Student Presentation

### Improving Metabolic Models Using Synthetic Lethality Data and Generating Genome-Scale Isotope Mapping Models for Flux Elucidation

Patrick F. Suthers<sup>1\*</sup> (suthers@enr.psu.edu), Alireza Zomorodi,<sup>1</sup> Prabhasa Ravikirthi,<sup>2\*</sup> and Costas D. Maranas<sup>1</sup>

<sup>1</sup>Dept. of Chemical Engineering, and <sup>2</sup>Dept. of Cell and Developmental Biology, Pennsylvania State University, University Park

<http://maranas.che.psu.edu>

**Project Goals:** The project goal of the research described here is twofold: First to improve the quality of genome-scale metabolic models by making use of gene essentiality and synthetic lethality experimental data. The second goal is to combine existing metabolic reconstructions with information from atom transitions to generate genome-scale isotope mapping models.

A pair of non-essential genes is referred to as synthetic lethal if the simultaneous deletion of both genes is lethal but the single gene deletions are not. One can generalize the concept of synthetic lethality to reactions or extend it further

by considering gene/reaction groups of increasing size where only the simultaneous elimination of all genes/reactions is lethal. Previous studies have demonstrated the utility of synthetic lethal predictions for the curation of genome-scale metabolic models. We recently used synthetic lethality information to identify twenty-one model improvements for the genome-scale model of *Escherichia coli*, *iAF1260*. In this talk, we discuss the systematic identification of synthetic lethal gene combinations for the most recent genome-scale metabolic model of yeast, (i.e., *iMM904*) for a variety of different growth medium conditions. By contrasting the *in silico* lethality predictions with *in vivo* observations we identified/corrected many missing regulatory mechanisms in yeast. The incorporation of the altered regulatory mechanisms into the genome-scale metabolic model led to a substantial increase in the accuracy of the *in silico* gene essentiality predictions. Overall, this study demonstrates the utility of synthetic lethality information for correcting genome-scale metabolic models.

Metabolic flux analysis (MFA) has so far been restricted to lumped networks lacking many important pathways, partly due to the difficulty in automatically generating isotope mapping matrices for genome-scale networks. Here we describe a procedure for the largely automated generation of atom mappings for genome-scale metabolic reconstructions. The developed procedure uses a compound matching algorithm based on the graph theoretical concept of pattern recognition along with relevant reaction heuristics to automatically generate genome-scale atom mappings which trace the path of atoms from reactants to products for every reaction in any given reconstruction. When applied to the *iAF1260* metabolic reconstruction of *Escherichia coli*, the genome-scale isotope mapping model *imPR90068* is obtained. The model maps 90,068 non-hydrogen atoms, contains  $1.37 \times 10^{157}$  distinct isotope forms and accounts for all 2,077 reactions present in *iAF1260* (the previous largest mapping model included 238 reactions). The expanded scope of *imPR90068* allows for tracking of labeled atoms through pathways such as cofactor and prosthetic group biosynthesis and histidine metabolism. We also discuss how using an elementary metabolite unit (EMU) representation of *imPR90068* significantly reduces the number of variables during MFA.

## 240

### Computational Pathway Identification and Strain Optimization for Chemical and Biofuel Production

Sridhar Ranganathan<sup>1\*</sup> (sur152@psu.edu), Patrick F. Suthers,<sup>2</sup> and **Costas D. Maranas**<sup>2</sup>

<sup>1</sup>Huck Institutes of the Life Sciences and <sup>2</sup>Dept. of Chemical Engineering, Pennsylvania State University, University Park

<http://maranas.che.psu.edu>

**Project Goals:** The main goal of this work is to develop new methods to discern novel pathways for chemical and

**biofuel production and to elucidate strain engineering strategies that will ensure production at desired target levels.**

We present an integrated computational base to support pathway identification and strain optimization with an emphasis on biofuel production. An efficient graph-based algorithm is presented for the exhaustive identification of all pathways enabling the production of a targeted biofuel molecule. The algorithm is based on a min-path formulation. It searches over a database of biotransformations that spans reactions from KEGG, Metacyc, BRENDA and other resources with an emphasis on C4+ alcohols. The identified pathways are then integrated into the genome-scale model of the production host (e.g., *Escherichia coli*). We describe the application of the OptForce computational framework to pinpoint engineering modifications (knock-outs/up/down) that are required for the targeted biofuel overproduction. This is accomplished by classifying reactions (and combinations thereof) in the metabolic model depending upon whether their flux values must increase, decrease or become equal to zero to meet the pre-specified overproduction target. A “force set” can then be extracted that contains a sufficient and non-redundant set of reactions that need to be directly changed to meet the production requirements. We apply the integrated framework for the production of 1-butanol, isobutanol, and other alcohols in *E. coli* using the most recent *in silico E. coli* model, *iAF1260*. We also examine the production of succinate in *E. coli*. The proposed computational workflow not only recapitulates existing pathways and engineering strategies but also reveals novel and non-intuitive ones that boost production by using and performing coordinated changes on sometimes distant pathways.

## 241

### COBRA Toolbox 2.0: *In Silico* Systems Biology Suite

Jan Schellenberger<sup>1\*</sup> (jschelle@ucsd.edu), Richard Que,<sup>2</sup> **Andrei Osterman**,<sup>3</sup> **Bernhard O. Palsson**,<sup>2</sup> and **Karsten Zengler**<sup>2</sup>

<sup>1</sup>Bioinformatics Program, and <sup>2</sup>Dept. of Bioengineering, University of California, San Diego; and <sup>3</sup>Burnham Institute for Medical Research, La Jolla, Calif.

[http://systemsbiology.ucsd.edu/Downloads/Cobra\\_Toolbox](http://systemsbiology.ucsd.edu/Downloads/Cobra_Toolbox)

**Project Goals:** This project is focused on a systems-level understanding of biological hydrogen production using *Thermotoga maritima* as a model organism. The project will address the basic science required to improve our understanding of hydrogen production from various carbon sources including glucose, cellulose, starch and xylan by this thermophilic microorganism. The overall goal is 1) to reconstruct the regulatory and metabolic network in *T. maritima* using various sets of “omics” data, 2) to integrate regulatory and metabolic networks into one “inte-

grated" genome-scale model, 3) to confirm and validate the ability of the integrated model to predict processing of various environmental signals.

With the advent of whole genome sequencing in the late 1990s, it became possible to build genome scale metabolic models. Since then, this field has undergone a renaissance in terms of 1) size and scope of reconstructions, 2) number of reconstructions and 3) number of analysis tools. The first version of the COBRA (Constraint Based Reconstruction and Analysis) toolbox was published in 2007 to combine many of these emerging methods into one easy to use package. We present version 2.0 here.

The COBRA toolbox is a set of Matlab scripts. Constraint Based models are loaded from various sources into a COBRA specific data structure. The user can then manipulate these models by using the command line or simple scripts. Methods can be chained to create simple data pipelines. The scope of COBRA falls under 8 categories as shown in Figure 1. New to version 2.0 are methods for gap filling, C13 analysis, visualization and thermodynamics. Also new in version 2.0 is a test case suite which gives examples of use of the different methods and expected results.

The objective of the COBRA toolbox is to abstract away details of implementation of constraint based methods. For the end user this reduces development time, cuts down on bugs and makes the code easier to share with other research groups. For the *Thermotoga* project, the COBRA toolbox is used to refine the model, analyze high throughput data, and visualize the results.

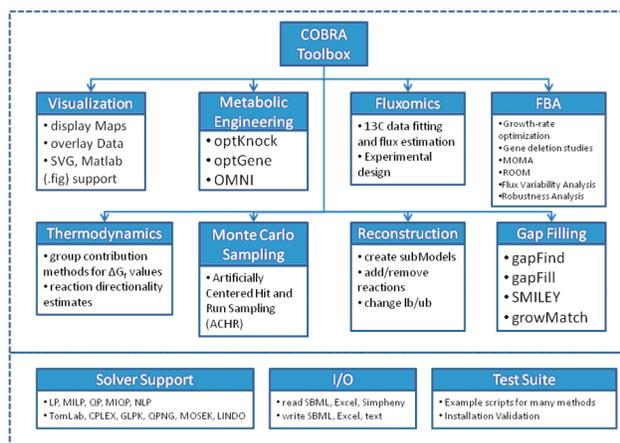


Figure 1: Features of the COBRA toolbox: Top) Scripts are available for methods in eight areas of metabolic systems biology. Bottom) Linear and Quadratic solvers are implemented through a simple yet flexible API in a vendor independent fashion. A set of test scripts are present to validate proper installation as well as demonstrate examples of use.

## Reference

1. Becker, S.A., Feist, A.M., Mo, M. L., Hannum, G., Palsson, B.Ø., Herrgard, M.J. Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox., *Nat. Protocols*, 2, 727-738 (2007).

# 242

## Numerical Optimization Algorithms and Software for Systems Biology: Optimality Principles in Nonequilibrium Biochemical Networks

Ronan M.T. Fleming<sup>1\*</sup> (ronan.mt.fleming@gmail.com), Chris Maes,<sup>2</sup> Ines Thiele,<sup>3</sup> Bernhard Ø. Palsson,<sup>4</sup> Yinyu Ye,<sup>5</sup> and Michael A. Saunders<sup>5</sup>

<sup>1</sup>Science Institute and Center for Systems Biology and <sup>2</sup>Center for Systems Biology, University of Iceland, Iceland; <sup>3</sup>Institute for Computational and Mathematical Engineering, Stanford University, Stanford, Calif.; <sup>4</sup>Dept. of Bioengineering, University of California, San Diego; and <sup>5</sup>Dept. of Management Science and Engineering, Stanford University, Stanford, Calif.

<http://www.hi.is/~rfleming>

### Project Goals: Simultaneous prediction of metabolic fluxes and concentrations in *Escherichia coli*.

We derive a new optimization problem on a steady-state non-equilibrium network of biochemical reactions, with the property that mass conservation, energy conservation, the second law of thermodynamics and the proportionality of reaction rate to reactant concentration, all hold at the problem solution. These nonlinear, non-convex constraints are enforced without recourse to linearization or any other form of approximation. This method provides the first computationally tractable method for enforcing thermodynamic, energy, and mass-conservation constraints, at genome scale. Moreover, the formalism has a clear thermodynamic interpretation and suggests a new optimality principle for non-equilibrium biochemical networks. This method may be used for simultaneously predicting reaction rate (flux) and metabolite concentrations in genome-scale biochemical networks. In particular, we demonstrate its utility for simultaneous integration of metabolomic and fluxomic data in *Escherichia coli*, in order to predict unmeasured concentrations and fluxes.