

GTL Program Projects

Harvard Medical School

Microbial Ecology, Proteogenomics and Computational Optima

A2

Microbial Ecology, Proteogenomics, and Computational Optima

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

Harvard Medical School and Massachusetts Institute of Technology

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

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

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

*Lawrence Berkeley National Laboratory*Rapid Deduction of Stress Response Pathways in Metal/
Radionuclide Reducing Bacteria**A4****Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria**

Adam Arkin^{2,3} (aarkin@lbl.gov), Alex Beliaev⁸, Inna Dubchak², Matthew Fields¹, Terry Hazen², Jay Keasling², Martin Keller⁴, Vincent Martin^{2,3}, Frank Olken², Anup Singh⁵, David Stahl⁷, Dorothea Thompson¹, Judy Wall⁶, and Jizhong Zhou¹

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

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

trol). To achieve this requires a more complete understanding of how the biological "units" comprising the system are organized, regulated, and linked in time and space (genes, genomes, cells, populations, communities, and ultimately, ecosystems). Key to these objectives is a more complete understanding of stress response systems and their environmental context.

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

Oak Ridge National Laboratory

Genomes to Life Center for Molecular and Cellular Systems

A Research Program for Identification and Characterization of Protein Complexes

A6**Bioinformatics and Computing in the Genomes to Life Center for Molecular and Cellular Systems**

D. A. Payne*¹ (debbie.payne@pnl.gov), E. S. Mendoza¹, G. A. Anderson¹, D. K. Gracio*¹, W. R. Cannon¹, T. P. Straatsma¹, H. J. Sofia¹, D. A. Dixon*¹, M. Shah², D. Xu², D. Schmoyer², S. Passovets², I. Vokler², J. Razumovskaya², T. Fridman², V. Olman², A. Gorin², E. Uberbacher², F. Larimer², and Y. Xu²

***Presenters**

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

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

LIMS System

We have purchased a commercially available and proven LIMS system, Nautilus™ (from Thermo Lab Systems) to serve as the backbone for integrating data management and analysis. Nautilus, once configured, will provide comprehensive sample tracking from planning through experimentation, data analysis, reporting, and final archival or disposal. Nautilus will be interfaced with labora-

tory instruments and data analysis tools and services to enable automation and standardization of data processing. Data will be archived through integration with the Environmental Molecular Sciences Laboratory Northwest File System archive.

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

Mass Spectrometry Proteomic Data Analysis

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

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

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

Bioinformatic Analysis Tools

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

This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute through Contract No. DE-AC06-76RLO 1830.

A8

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

Gregory B. Hurst¹ (hurstgb@ornl.gov), Robert L. Hettich¹, Nathan C. Verberkmoes¹, Gary J. Van Berkel¹, Frank W. Larimer¹, Trish K. Lankford¹, Steven J. Kennel¹, Dale Pelletier¹, Jane Razumovskaya¹, Richard D. Smith², Mary Lipton², Michael Giddings⁵, Ray Gesteland⁴, Malin Young³, and Carol Giometti⁶

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

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

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

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

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

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

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

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

A10

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

Joshua N. Adkins¹, Deanna Auberry¹, Baowei Chen¹, James R. Coleman¹, Priscilla A. Garza¹, Jane M. Weaver Feldhaus¹, Michael J. Feldhaus¹, Yuri A. Gorby¹, Eric A. Hill¹, Brian S. Hooker¹, Chian-Tso Lin¹, Mary S. Lipton¹, L. Meng Markillie¹, M. Uljana Mayer¹, Keith D. Miller¹, Sewite Negash¹, Margaret F. Romine¹, Liang Shi¹, Robert W. Siegel¹, Richard D. Smith¹, David L. Springer¹, Thomas C. Squier¹, **H. Steven Wiley¹** (steven.wiley@pnl.gov), Linda J. Foote², Trish K. Lankford², Frank W. Larimer², T-Y. S. Lu², Dale Pelletier², Stephen J. Kennel², and Yisong Wang²

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

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

Cloning, Expression, and Purification: To date, 23 *S. oneidensis* genes have been cloned into the GATEWAY™ expression vector pDEST™ containing a His₆-tag for purification. Initial screening tests indicate that ~73% of cloned genes were expressed. Among those expressed proteins, 8 were purified to homogeneity using a Ni-NTA column under nondenaturing conditions. The yields of purified proteins obtained from 1 L of culture varied from 5 to 29 mg. We have also constructed new GATEWAY™-compatible vectors that will permit the expression of His₆-tagged proteins in both *S. oneidensis* and *R. palustris* and the subsequent isolation of preformed complexes from microbes. Using four modified pDEST vectors, 7 *R. palustris* genes have been cloned and expressed in *E. coli*. We are testing both N and C-terminal 6-his and GST tags for efficiency of expression and purification. Western blots of proteins and MS spectra of tryptic digests (see MS poster) of the GST-tagged nitrite reductase verify the expression and purification of polyproteins at high yield. The modified vector containing the GroEL gene has been inserted into *R. palustris* and it appears to be retained and convey drug resistance to the bacteria. Pull down experiments are in progress to isolate complexes from this target organism.

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

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

conditions a limited range of cross-linked products are observed that are readily characterized by mass spectrometry.

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

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

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

This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute through Contract No. DE-AC06-76RLO 1830. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U. S. Department of Energy under Contract No. DE-AC05-00OR22725.

A12

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

Michelle Buchanan¹ (buchananmv@ornl.gov), Frank Larimer¹, Steven Wiley², Steven Kennel¹, Thomas Squier², Michael Ramsey¹, Karin Rodland², Gregory Hurst¹, Richard Smith², Ying Xu¹, David Dixon², Mitchel Doktycz¹, Steve Colson², Carol Giometti³, Raymond Gesteland⁴, Malin Young⁵, and Michael Giddings⁶

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

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

networks and molecular level characterization of interactions in microbial communities.

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

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

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

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

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

and dissemination, along with computational tools for modeling and simulation of protein complexes are being developed.

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

A14

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

P. R. Hoyt¹ (hoytpr@ornl.gov), C. J. Bruckner-Lea², S. J. Kennel¹, P. K. Lankford¹, M. S. Lipton², R. S. Foote¹, J. M. Ramsey¹, K. D. Rodland², and M. J. Doktycz¹

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

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

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

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

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

An heterologous-tagged protein pulldown system, for *S. oneidensis* using single-chain antibodies (Ab) to specific expressed proteins is also under automation development. This process uses a microfluidics platform combined with functionalized microbeads for the purification of protein complexes. A renewable microcolumn system with optical detection has been assembled and automated procedures developed. The renewable microcolumn consists of small volumes (microliters) of microbeads that are automatically packed, perfused with cell lysates, and wash solutions, and proteins eluted using a solution that is suitable for mass spectrometry analysis. After each purification, the small volume of microbeads is automatically flushed from the microcolumn and a new microcolumn is automatically packed. The microbeads are functionalized for the capture of a specific protein, for example by derivatization

with an antibody for the protein of interest. Optical monitoring of the microcolumn during processing provides information about the amount of material on the column during each binding and washing step. The current automated procedure can process a cell lysate volume ranging from 10 microliters to 1 milliliter, and the purified proteins are eluted into 150 microliters of a low salt buffer solution. Automated procedures are currently being tested for the capture of *Shewanella* proteins tagged with yellow fluorescent protein (YFP), along with the proteins that associate with the YFP-tagged protein. As new reagents for protein capture such as single chain antibodies for *Shewanella* proteins of interest are developed, they will be linked to microbeads and renewable column protocols will be developed for automated purification of the protein complexes for mass spectrometry. In the next stage of this work, the eluted protein complexes will be analyzed by mass spectrometry and the automated protocols will be optimized.

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

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

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

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

Sandia National Laboratories

Carbon Sequestration in *Synechococcus*

From Molecular Machines to Hierarchical Modeling

A16

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

Anthony Martino¹ (martino@sandia.gov), Andrey Gorin², Todd Lane¹, Steven Plimpton¹, Nagiza Samatova², Ying Xu², Hashim Al-Hashimi³, Charlie Strauss⁴, Byung-Hoon Park², George Ostrouchov², Al Geist², William Hart², and Diana Roe¹

¹Sandia National Laboratories, P.O. Box 969, MS9951, Livermore, CA 94551; ²Oak Ridge National Laboratory, P.O. Box 2008, MS6367, Oak Ridge, TN 37831;

³University of Michigan, Department of Chemistry, 930 N. University, Ann Arbor, MI 48109; and ⁴Los Alamos National Laboratories, P.O. Box 1663, Los Alamos, NM 87545

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

Experimental Elucidation of Protein Complexes

Initial efforts will focus on the carboxysome, a polyhedral inclusion body that consists of a protein shell surrounding ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO). While RuBisCO regulates photosynthetic carbon reduction, the function of the carboxysome is unclear. The carboxysome may either actively promote carbon fixation by concentrating CO₂ or passively play a role by regulating RuBisCO turnover. The presence of carbonic anhydrase, an enzyme that regulates the equilibrium between inorganic car-

bon species, in the carboxysome would suggest an active role in carbon concentration, but experimental results are mixed. No clear biochemical evidence for a link between carbonic anhydrase and the carboxysome exists in WH8102.

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

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

Computational Elucidation of Protein Complexes

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

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

In addition to “sequence-based” protein signatures one of our main aims is the development of structure-based algorithms for the inference of protein-protein interactions. At the initial stage we will apply structure prediction methods to determine protein fold families with our ROSETTA and PROSPECT programs and use inferred structural similarities to create hypotheses

about their interacting partners. The necessary step in this process is a creation of structure prediction pipeline for high throughput characterization of the protein folds. The computational pipeline merges several bioinformatics and modeling tools including algorithms for protein domain division, secondary structure prediction, fragment library assembly, and structure comparison. Since the protein folding algorithms deliver not unique answers but rather ensembles of predictions we will also construct database system to store and curate the accumulated inferences.

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

A18

Carbon Sequestration in *Synechococcus*: Microarray Approaches

Brian Palenik⁴, Anthony Martino², **Jerilyn A. Timlin**² (jatimli@sandia.gov), David M. Haaland², Michael B. Sinclair², Edward V. Thomas², Vijaya Natarajan³, Arie Shoshani³, Ying Xu¹, Dong Xu¹, Phuongan Dam¹, Bianca Brahamsha⁴, Eric Allen⁴, and Ian Paulsen⁵

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

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

A20

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

Grant S. Heffelfinger¹ (gsheffe@sandia.gov), Anthony Martino², Andrey Gorin³, Ying Xu³, Mark D. Rintoul III¹, Al Geist³, Hashim M. Al-Hashimi⁸, George S. Davidson¹, Jean Loup Faulon¹, Laurie J. Frink¹, David M. Haaland¹, William E. Hart¹, Erik Jakobsson⁷, Todd Lane², Ming Li⁹, Phil Locascio², Frank Olken⁴, Victor Olman², Brian Palenik⁶, Steven J. Plimpton¹, Diana C. Roe², Nagiza F. Samatova³, Manesh Shah², Arie Shoshani⁴, Charlie E. M. Strauss⁵, Edward V. Thomas¹, Jerilyn A. Timlin¹, and Dong Xu²

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

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

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

A22

Systems Biology Models for *Synechococcus sp.*

Mark D. Rintoul¹ (rintoul@sandia.gov), Damian Gessler², Jean-Loup Faulon¹, Shawn Means¹, Steve Plimpton¹, Tony Martino², and Ying Xu³

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

Ultimately, all of the data that is generated from experiment must be interpreted in the context of a model system. Individual measurements can be related to a very specific pathway within a cell, but the real goal is a systems understanding of the cell. Given the complexity and volume of experimental data as well as the physical and chemical models that can be brought to bear on subcellular processes, systems biology or cell models hold the best hope for relating a large and varied number of measurements to explain and predict cellular response. Clearly, cells fit the working scientific definition of a complex system: a system where a number of simple parts combine to form a larger system whose behavior is much harder to understand. The primary goal of this subproject is to integrate the genomic data generated from the overall project's experiments and lower level simu-

lations, along with data from the existing body of literature, into a whole cell model that captures the interactions between all of the individual parts. It is important to note here that all of the information that is obtained from other efforts in this project is vital to the work here. In a sense, this is the "Life" of the "Genomes to Life" theme of this project.

The precise mechanism of carbon sequestration in *Synechococcus sp.* is poorly understood. There is much unknown about the complicated pathway by which inorganic carbon is transferred into the cytoplasm and then converted to organic carbon. While work has been carried out on many of the individual steps of this process, the finer points are lacking, as is an understanding of the relationships between the different steps and processes. Understanding the response of *Synechococcus sp.* to different levels of CO₂ in the atmosphere will require a detailed understanding of how the carbon concentrating mechanisms in *Synechococcus sp.* work together. This will require looking these pathways as a system.

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

University of Massachusetts, Amherst

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

A24

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

Derek Lovley¹ (dlovley@microbio.umass.edu), Stacy Ciufu¹, Zhenya Shebolina¹, Abraham Esteve-Nunez¹, Cinthia Nunez¹, Richard Glaven¹, Regina Tarallo¹, Daniel Bond¹, Maddalena Coppi¹, Pablo Pomposiello¹, Steve Sandler¹, Barbara Methé², Carol Giometti³, and Julia Krushkal⁴

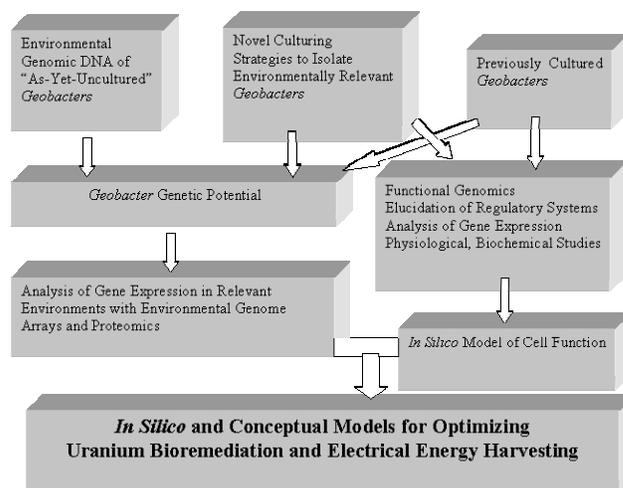
¹University of Massachusetts; ²The Institute for Genomic Research; ³Argonne National Laboratory; and ⁴University of Tennessee

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

Progress to Date: Although the physiology of pure cultures of *Geobacters* are being studied and modeled in detail, the degree of similarity in the genetic potential of the *Geobacters* in culture and those that predominate during uranium bioremediation or electrical energy harvesting is unknown. The environmental component of the studies in the first four months of this project have focused a NABIR-program site, located in Rifle, Colorado in which the addition of acetate to the subsurface stimulated the growth of *Geobacter* species and the removal of uranium from the groundwater. In order to evaluate the genetic potential of the *Geobacter* species involved

in uranium bioremediation, which at times accounted for over 80% of the total microbial community in the groundwater, genomic DNA was extracted from sediments undergoing active uranium reduction and is now being sequenced at the Joint Genome Institute. Some of this data should be available by the time of the meeting and novel methods for assembling complete or nearly complete *Geobacter* genomes from this environmental genomic DNA will be presented. An additional strategy to learning more about the genetic potential of *Geobacters* living in the subsurface was to isolate the predominant *Geobacters* and sequence their genomes. Using a novel technique, we were able to isolate a *Geobacter* from the study site whose 16S rDNA sequenced matched a 16S rDNA sequence that was prevalent in clone libraries from the uranium reduction zone at the study site. The genome of this organism will be studied in detail in the next year.

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



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

as a repressor signal controlling the expression of the fumarate reductase genes will also be described.

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