GTL Milestone 1

Develop Techniques to Determine the Genome Structure and Functional Potential of Microbes and Microbial Communities

Section 1

Organism Sequencing, Annotation, and Comparative Genomics

1 GTL

Genomic Reconstruction and Experimental Validation of Catabolic Pathways in *Shewanella* Species

Andrei Osterman[1,2]* (osterman@burnham.org), Dmitry Rodionov,[1] Chen Yang,[1] Yanbing Wang,[4] Margaret Romine,[3] Anna Obraztsova,[4]* and **Kenneth Nealson[4]**

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

Project Goals: This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics, leveraging the availability of genome sequence for 18 additional strains of *Shewanella*, to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus. To understand these systems the SF is using genome-based approaches to investigate *Shewanella* as a system of integrated networks; first describing key cellular subsystems— those involved in signal transduction, regulation, and metabolism—then building towards understanding the function of whole cells and, eventually, cells within populations. As a general approach, the SF is collectively employing complimentary "top-down"—bioinformatics-based genome functional predictions, high-throughput expression analyses, and functional genomics approaches to uncover key genes as well as metabolic and regulatory networks. The "bottom-up" component employs more traditional approaches including genetics, physiology and biochemistry to test or verify predictions. This information will ultimately be linked to analyses of signal transduction and transcriptional regulatory systems and used to develop a linked model that will contribute to understanding the ecophysiology of *Shewanella* in redox stratified environments.

Knowledge of the pathways and mechanisms of carbon assimilation and utilization for biomass and energy production is one of the key aspects of our understanding of environmental microorganisms in the context of respective ecosystems. Comparative genomic analysis has revolutionized our ability to quickly predict which metabolic subsystems occur in newly sequenced genomes, the set of genes of which each is comprised, and to suggest their functional roles within each subsystem. Results of this predictive analysis can then be used to design and conduct targeted physiological and biochemical assays to validate novel conjectures of gene function revealed in this process. By taking advantage of such computational predictions one can dramatically reduce the volume of experimental studies required to assess basic metabolic properties of multiple bacterial species.

The current availability of partial or complete genome sequences for 16 *Shewanella* strains provides and unprecedented opportunity for a systematic comparative analysis of this important group of species. For example, the integrative genomic approach was successfully used in our recent analysis of the novel N-acetylglucosamine (GlcNAc) utilization subsystem in *S. oneidensis* and related species. By using subsystem reconstruction and genome context analysis tools provided in The SEED genomic platform (http://theseed.uchicago.edu/FIG/index.cgi) we tentatively assigned a number of novel genes including GlcNAc-related transporter (NagP, SO3503), transcriptional regulator (NagR, SO3516) and enzymes, GlcNAc kinase (SO3507) and GlcN-6P deaminase (SO3506) that are non-homologous to the respective components of *E. coli* GlcNAc pathway. Two latter assignments and the whole biochemical pathway of GlcNAc conversion to Fructose-6P were experimentally verified by in vitro reconstitution. The results of phenotypic profiling were fully consistent with genomic reconstruction as only one strain, *S. frigidimarina*, which lacked the respective genes, was unable to grow on GlcNAc.

Extension of this approach to the analysis of a large panel of catabolic pathways has revealed substantial differences between *Shewanellae* and well-studied model species, such as *E. coli*. These differences occur at the level of presence or absence of the entire pathways, the use of alternative biochemical routes, different regulatory mechanisms and nonorthologous displacements of individual genes. For example, in contrast to *E. coli*, all of the sequenced *Shewanella* ssp. possess the elaborate machinery for catabolism of branch chain amino acids and histidine. Under experimental conditions used so far, most (but not all) of the tested *Shewanellae* were unable to grow on these amino acids as sole carbon sources. Further experiments are currently under way to reconcile these apparent inconsistencies. At the same time, both bioinformatics and experimental analyses revealed a fully consistent distribution of glycerate and sucrose utilization pathways among different strains of *Shewanella*. The genomic reconstruction of both pathways (not present in *E. coli*) allowed U.S. to predict novel gene families, including putative glycerate transporter (e.g., SO1771) as well as regulatory and uptake components of sucrose metabolism (in the *scr* operon, Sfri1095-1099, of *S. frigidimarina*). Using a genome context, such as clustering on the chromosome and long-range homology analysis, we were able to predict candidate genes for various functional roles in other pathways (e.g., fatty acid metabolism). In addition to the physiological analysis of the *Shewanellae*, we are currently pursuing experimental validation of several functional predictions by a variety of biochemical and genetic techniques. Results of these analyses are an integral component of ongoing research of ***Shewanella* Federation** whose ultimate goal is to develop a better understanding of *Shewanella* ecophysiology and speciation.

2 GTL

Evolutionary Analysis of Proteins Deduced from 10 Fully Sequenced *Shewanella* Genomes

N. Maltsev[1]* (maltsev@mcs.anl.gov), D. Sulakhe,[1] A. Rodriguez,[1] M. Syed,[1] and M. Romine[2]

[1]Argonne National Laboratory, Argonne, Illinois and [2]Pacific Northwest National Laboratory, Richland, Washington

Project Goals: The proteomes of the 10 completed *Shewanella* genomes (total 43,839 protein sequences) were analyzed using GNARE (GeNome Analysis and Research Environment). This Grid-based computational system for automated high-throughput analysis of genomes and metabolic reconstructions is being developed by Bioinformatics group at MCS, Argonne National

* Presenting author

Laboratory. In addition to offering users with tools for annotating protein functions, GNARE provides automated metabolic reconstructions from the sequence data to facilitate identification of missing enzymes and comparative analysis of metabolic pathways. GNARE allows comparison of metabolic models of *Shewanella* strains with over 300 metabolic models for completely sequenced prokaryotic organisms. Chisel, a workbench for evolutionary analysis of enzymes being developed by our group, was used as a supporting tool for automated prediction of function, identification of taxonomy-specific metabolic signatures and identification of cases of potential horizontal gene transfer. The poster will present a detailed examination of these analyses.

The Joint Genome Institute has produced closed genome sequences for nine *Shewanella* strains. The environments from which these 9 strains and the first strain sequenced, *S. oneidensis* MR-1, were collected from vary from fresh and marine waters and underlying sediments to terrestrial sediments. The availability of this diverse set of genomes provides a unique opportunity to explore protein and metabolic pathway evolution within a single Genus. We will provide an overview of the unique suite of computational tools that we have begun applying to this set of sequences for both the purposes of updating the annotation and for studying cellular evolution.

Genome Annotation Using GNARE. The proteins deduced from the 10 completed *Shewanella* genomes (total 43,839 protein sequences) were analyzed using GNARE (GeNome Analysis and Research Environment)—a Grid-based computational system for automated high-throughput sequence characterization and metabolic reconstructions. This system uses conventional bioinformatics tools including BLAST, Blocks, InterPro as well as specialty tools developed by our group, including Chisel and a function prediction voting algorithm. The analysis was performed on 1317 CPUs of distributed computational resources from the Open Science Grid and TeraGrid allowing analysis of all 10 *Shewanella* proteomes to be completed in 64 hours. The process yields a suite of web pages that enables the *Shewanella* federation annotation team to view the results of the automated analysis and that facilitates further searches for evidence of function via a broad variety of additional computational tools provided by Puma2. Because the system was specifically designed for cross-genome comparative analysis it is well suited for comparative analysis of predicted functions associated with the various *Shewanella* proteomes.

Metabolic reconstructions. In addition to providing users with tools for annotating protein functions, GNARE provides automated metabolic reconstructions (accessible via a web interface) for each genome that can be viewed via either KEGG or EMP maps to facilitate identification of missing enzymes and subsequent searches for candidate proteins to fill these metabolic pathway gaps through use of high-resolution tools, such as Dragonfly and PhyloBlocks, developed by our group. As annotations are updated in the GNARE system they can readily be propagated through the metabolic constructions, thereby further facilitating the process of predicting protein functions in context of metabolic pathways. Simultaneous analysis of all 10 proteomes provides a foundation for comparative analysis of metabolic pathways characteristic of particular strains. We will present examples of the resulting predicted metabolic diversity among the various strains of *Shewanella*.

Identification of taxonomy-specific metabolic signatures using Chisel. A workbench for identification and evolutionary analysis of taxonomic and phenotypic variations of enzymes Chisel was used as a supporting tool for automated prediction of function for the 10 *Shewanella* proteomes. Chisel predicted enzyme assignments for 9,946 of a total 43,691 proteins in the 10 genomes. The number of EC assignments for individual genomes ranged from 718 in *S. denitrificans* OS217 to 1,081 for *S. oneidensis* MR-1. Interestingly, out of 1,081 enzymes predicted by Chisel in MR-1, 598 proteins corresponded to enzyme variations specific to *Proteobacteria*; 217 proteins to *Gammaproteobacteria*; 11 to *Alteromonadales* and 39 enzymes to the *Shewanella* Genus. Such variations in levels of taxonomic specificity indicate that enzymes in metabolic pathways evolve at different rates. In the course of

adaptation some of the enzymes become more specific to particular taxonomies. The *Shewanella*-specific variations of enzymes were found to be associated predominantly with core metabolic pathways (e.g., glycolysis, purine and pyrimidine biosynthesis, metabolism of amino acids) as well as chemotaxis and sensory transduction processes. This observation suggests significant systems-level adaptation that led to diversification of enzymes in this group of organisms in the course of evolution. Identification of taxonomy-specific signature enzymes may provide insights into mechanisms driving the emergence of taxonomy and phenotype-specific pathways. The Chisel system also supports the development of PCR primers and oligonucleotides corresponding to these models using the CODEHOP program (Henikoff and Henikoff, 1996). This feature can assist experimentalists in identifying particular enzymatic functions in organisms of interest using biochip- or PCR-based technologies.

Discovery of potential cases of horizontal gene transfer. The Chisel analysis also helps to identify potential cases of horizontal gene transfer. According to our analysis a significant number of enzymatic genes in *Shewanella* appear to have been acquired from *Cyanobacteria* and *Firmicutes*. For example, analysis of the MR-1 revealed 18 proteins most similar to Cyanobacterial enzymes and 36 enzymes that are most closely related to various *Firmicutes*.

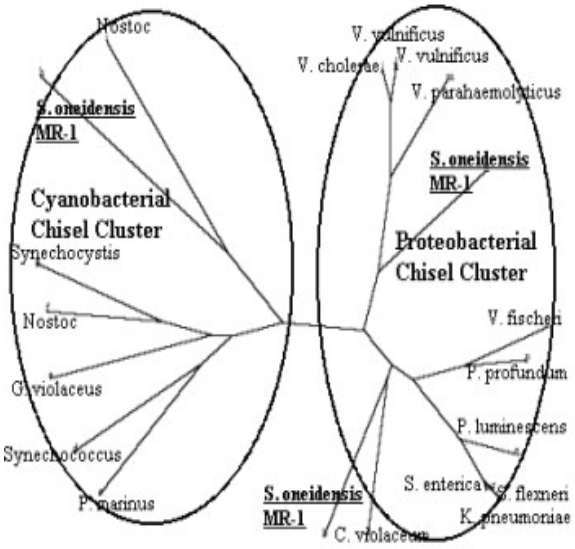


Fig 2. Phylogenetic tree for peptide deformylase (EC 3.5.1.88). MR-1 has 3 copies of this enzyme: 2 were classified by Chisel algorithm to a proteobacterial cluster (SF004749_6_B_Proteobacteri8) and one to a cyanobacterial cluster (SF004749_6_B_Cyanobacteria4).

The poster will present a detailed examination of these findings.

* Presenting author

3 GTL

Modeling Conserved Indels as Phylogenetic Markers in *Shewanella*

John P. McCrow[1]* (mccrow@usc.edu), **Kenneth H. Nealson**,[2] and **Michael S. Waterman**[1]

[1]Computational Molecular Biology, University of Southern California, Los Angeles, California and [2]Geobiology, University of Southern California, Los Angeles, California

Project Goals: This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics, leveraging the availability of genome sequence for 18 additional strains of *Shewanella*, to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus. To understand these systems the SF is using genome-based approaches to investigate *Shewanella* as a system of integrated networks; first describing key cellular subsystems those involved in signal transduction, regulation, and metabolism—then building towards understanding the function of whole cells and, eventually, cells within populations. As a general approach, the SF is collectively employing complimentary "top-down" bioinformatics-based genome functional predictions, high-throughput expression analyses, and functional genomics approaches to uncover key genes as well as metabolic and regulatory networks. The "bottom-up" component employs more traditional approaches including genetics, physiology and biochemistry to test or verify predictions. This information will ultimately be linked to analyses of signal transduction and transcriptional regulatory systems and used to develop a linked model that will contribute to understanding the ecophysiology of *Shewanella* in redox stratified environments.

Phylogenetic signal derived from small subunit rRNA or core protein sequences is especially confusing and noisy among bacterial genomes, perhaps partially due to elevated levels of lateral gene transfer between lineages. The use of Rare Genomic Changes (RGCs) as phylogenetic characters has been proposed as potentially yielding higher quality information about the evolutionary history of genomes than collections of nucleotide or amino acid substitutions. High level genomic information including gene content, gene order, conserved inserts or deletions (indels), and nucleotide composition such as GC content, have been used to track evolutionary relationships between organisms within common core gene sets. However, the use of RGCs is limited in mainstream phylogenetic analysis because, unlike for conserved amino acid sequences, no statistical models exist to describe these high level genomic events.

We describe a method for the direct alignment, statistical modeling, and integration of conserved indels with adjacent amino acid sequence for improved phylogenetic signal. The best multiple alignment of amino acid sequences is not suitable for directly aligning conserved indels. Instead, we perform all pairwise alignments around potential indels, and filter out those with ambiguous classifications, as far as conserved indel states, to reduce homoplasy. The resulting set of conserved indels defines the highest confidence splits and are used as phylogenetic characters to infer the historical branching order of the species involved. While the extent of homoplasy within conserved indels is assumed to be low, we know of no previous statistical models or attempts to quantify independent convergent events or reversion rates of conserved indels. By integrating adjacent amino acid sequences with their corresponding conserved indels we can estimate the extent of homoplasy within conserved indels themselves, as well as improve the accuracy of phylogenetic signal from conserved proteins.

Conserved indels have great potential in aiding our understanding and validation of both the historical branching order of species, as well as the identification of lateral gene transfer events. They may

also be particularly suited for use as simple molecular markers of lineage and have the potential to be used to easily identify species present in an unknown mixed or pure sample without the need for any sequencing. Here we focus on exploring the utility of conserved indels to describe the phylogenetic relationships among sequenced species of the bacterial genus *Shewanella*.

4 GTL

Shewanella Population Comparative Genomics and Proteomics: Connecting Speciation, Ecophysiology, and Evolution

Jorge L.M. Rodrigues[1]* (rodrig76@msu.edu), Konstantinos T. Kostantinidis,[2] Margaret F. Romine,[3] Margrethe H. Serres,[4] Lee Ann McCue,[3] Mary S. Lipton,[3] Carol S. Giometti,[4] Anna Obraztova,[5] Matt Marshall,[3] Miriam Land,[6] Kenneth H. Nealson,[5] James K. Fredrickson,[2] and **James M. Tiedje**[1]

[1]Michigan State University, East Lansing, Michigan; [2]Massachusetts Institute of Technology, Cambridge, Massachusetts; [3]Pacific Northwest National Laboratory, Richland, Washington; [4]Marine Biological Laboratory, Woods Hole, Massachusetts; [4]Argonne National Laboratory, Argonne, Illinois; [5]University of Southern California, Los Angeles, California; and [6]Oak Ridge National Laboratory, Oak Ridge, Tennessee

Project Goals: Integrated genome-based studies of *Shewanella* ecophysiology. The overall goal of this project is to apply the tools of genomics, to better understand the ecophysiology and speciation of respiratory-versatile members of the *Shewanella* genus.

Shewanella is a very versatile microorganism, capable of respiring more than 10 electron acceptors and found in many different environments. Therefore, *Shewanella* has potential to be used in different bioremediation schemes ranging from nitrate contaminant removal to metal and radionuclides reduction/immobilization processes. Previous molecular and physiological studies have primarily focused on *Shewanella oneidensis* MR-1, but it is still unclear whether prokaryotic systems biology can rely on a single model organism to predict functional responses for its entire population. Here, we extend the knowledge about the *Shewanella* genus by comparing closed genomes and their respective proteomes of 10 strains of this genus, including: *S. oneidensis* MR-1, *S. putrefaciens* W3-18-1, *S. putrefaciens* CN-32, *Shewanella* sp. MR-7, *Shewanella* sp. MR-4, *Shewanella* sp. ANA-3, *S. denitrificans* OS217, *S. frigidimarina* NCIMB 400, *S. loihica* PV-4, and *S. amazonensis* SB2B. The availability of these genomes allows questions towards the following aims: 1) to identify the gene core content of *Shewanella* genus, 2) to find genetic differences responsible for the ecological and physiological differentiation among the strains, and 3) to identify the ecological forces being implemented at genome level leading to speciation.

Results from comparative genomics showed that the above strains vary from 70 to 98.4% on pairwise average nucleotide identity (ANI). These results reveal a continuum of genetic relatedness for all sequenced genomes. The gene core dataset was calculated with use of three different methods giving similar values: 1817 for reciprocal DNA best match (Konstantinidis and Tiedje (2005), 1984 for protein alignments of 70% and a scoring matrix with Pam value of 100 (Serres and Riley 2006), and 2075 for pairwise reciprocal BLAST. The gene core dataset identified in all three sets is 1718 genes. Genome synteny deteriorated rapidly as the ANI decreased to values below 80%, indicating extensive chromosomal rearrangements that might have significant functional impact on the phenotypic and proteomic profiles of *Shewanella* species. The predicted central metabolism is almost identical and we have identified over 90 conserved pathways to this date for all sequenced genomes. A survey of

* Presenting author

unique genes belonging to each of the strains revealed that the majority fell into select categories: 1) hypotheticals, 2) mobile elements, 3) motility and attachment, 4) sensory, and 5) regulatory genes. Co-localization of many of the unique genes on the genome suggests that many may have been acquired via lateral transfer. These differences might indicate that ecological forces are being implemented at the genome level, allowing short term niche adaptation (for closely related strains), leading to later speciation (distantly related species).

Proteomic analyses using two-dimensional electrophoresis and ion trap mass spectrometry were performed for all sequenced strains, resulting in larger differences at proteomic level in comparison to genomic analyses. These results might indicate extensive differences at the regulatory level, since all strains were grown under identical conditions.

Shewanella denitrificans is the only member of this group unable to reduce iron and hence provides an excellent tool to investigate which genes may be responsible for adaptation to this metabolic resource. Functions unique to the metal reduced that were highlighted by this type of comparative analysis include: 1) quinol:fumarate reductase, 2) Fe(III) permease, 3, glycogen metabolism, 4) lactate oxidation, 5) NiFe hydrogenase, and 6) a large cluster of fatty acid biosynthetic genes. This strain is devoid of any proteins containing more than four cytochrome c heme binding motifs or of metaquinone biosynthetic genes. While also found in *S. denitrificans* an alternative variant of Na-translocating NADH-quinone reductase and ammonifying nitrate reductase was found only in the metal reducting *Shewanellae. S. denitrificans* lacks various signal transduction/regulatory proteins and transporters that are believed to be associated with anaerobic metabolism.

These results highlight the power of comparative bioinformatics, proteomics, and phenotypic analyses of related sequenced strains to acquire a greater understanding of evolution and ecophysiology speciation.

References

1. Kostantinidis, K. and J.M. Tiedje. 2005. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. U.S.A. **7:**2567-2572.
2. Serres, M.H. and M. Riley. 2006. 2006. Genomic analysis of carbon source metabolism of *Shewanella oneidensis* MR-1: predictions versus experiments. J. Bacteriol. **188:**4601-4609.

5 GTL

The Complete Genome of the Uncultivated Ultra-Deep Subsurface Bacterium *Desulforudis audaxviator* Obtained by Environmental Genomics

Dylan Chivian[1,2]* (DCChivian@lbl.gov), Eric J. Alm,[1,3] Eoin L. Brodie,[2] David E. Culley,[4] Thomas M. Gihring,[5] Alla Lapidus,[6] Li-Hung Lin,[7] Steve Lowry,[6] Duane P. Moser,[8] Paul Richardson,[7] Gordon Southam,[9] Greg Wanger,[9] Lisa M. Pratt,[10] **Adam P. Arkin**[1,2,11,12,13] (aparkin@lbl.gov), Terry C. Hazen,[1,2] Fred J. Brockman,[4] and Tullis C. Onstott[14]

[1]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov; [2]Lawrence Berkeley National Laboratory, Berkeley, California; [3]Massachusetts Institute of Technology, Cambridge, Massachusetts; [4]Pacific Northwest National Laboratory, Richland, Washington; [5]Florida State University, Tallahassee, Florida; [6]DOE Joint Genome Institute, Berkeley, California; [7]National Taiwan University, Taipei, Taiwan; [8]Desert Research Institute, Las Vegas, Nevada; [9]University of Waterloo, London, Ontario, Canada; [10]Indiana University, Bloomington, Indiana; [11]University of California, Berkeley, California; [12]Howard Hughes Medical Institute, Chevy Chase, Maryland; [13]Department of Bioengineering, University of California, Berkeley, California; and [14]Princeton University, Princeton, New Jersey

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

A more complete picture of life on Earth, and even life *in* the Earth, has recently become possible through the application of environmental genomics. We have obtained the complete genome sequence of a new genus of the *Firmicutes*, the uncultivated sulfate reducing bacterium *Desulforudis audaxviator*, by filtering fracture water from a borehole at 2.8 km depth in a South African gold mine. The DNA was sequenced at the JGI using a combination of traditional Sanger sequencing and 454 pyrosequencing, and assembled into just one genome, indicating the planktonic community is extremely low in diversity. We analyzed the genome of *D. audaxviator* using the MicrobesOnline annotation pipeline and toolkit (http://www.microbesonline.org, and see MicrobesOnline abstract), which offers powerful resources for comparative genome analysis, including operon predictions and tree-based comparative genome browsing. MicrobesOnline allowed U.S. to compare the *D. audaxviator* genome with other sequenced members of the *Firmicutes* in the same clade (primarily *Pelotomaculum thermoproprionicum*, *Desulfotomaculum reducens*, *Carboxydothermus hydrogenoformans*, and *Thermoanaerobacter tengcongensis*), as well as other known sulfate reducers (including *Archaeoglobus fulgidus* and *Desulfovibrio vulgaris*). *D. audaxviator* gives a view to the set of tools necessary for what appears to be a self-contained, independent lifestyle deep in the Earth's crust. The genome is not very streamlined, and indicates a motile, endospore forming sulfate reducer with pili that can fix its own nitrogen and carbon. *D. audaxviator* is an obligate anaerobe, and lacks obvious homologs of many of the traditional O_2 tolerance genes, consistent with the low concentration of O_2 in the fracture water and its long-term isolation from the surface. *D. audaxviator* provides a complete genome representa-

* Presenting author

tive of the Gram-positive bacteria to further our understanding of dissimilatory sulfate reducing bacteria and archaea, and offers the full complement of genes necessary for an independent lifestyle based solely on interactions with the geochemistry of the deep subsurface.

6 GTL

Genomic Comparisons Between a Metal-Resistant Strain of *Desulfovibrio vulgaris* and the Type Strain *D. vulgaris* Hildenborough

C.B. Walker,[1,4] D. Joyner,[2,4] D. Chivian,[2,4] S.S. Stolyar,[1,4] K. Hillesland,[1,4] J. Gabster,[1,4] P. Dehal,[2,4] M. Price,[2,4] T.C. Hazen,[2,4] **A.P. Arkin**[2,4] (aparkin@lbl.gov), P.M. Richardson,[3] D. Bruce,[3] and D.A. Stahl[1,4]*

[1]University of Washington, Seattle, Washington; [2]Lawrence Berkeley National Laboratory, Berkeley, California; [3]DOE Joint Genome Institute, Walnut Creek, California; and [4]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

As part of the Virtual Institute for Microbial Stress and Survival (VIMSS), the Environmental Stress Pathway Project (ESPP) investigates the metabolic response of a sulfate-reducing bacterium, *Desulfovibrio vulgaris* Hildenborough, to contaminants found at DOE sites. Under this aegis, the ESPP Applied Environmental Core (AEC) seeks to isolate and characterize environmentally relevant sulfate-reducing bacterium from contaminated sites. Comparative analyses between these isolates and *Desulfovibrio vulgaris* Hildenborough provides an informative framework for further elucidating mechanisms of microbial adaptation to environmental stressors. To this end, a sulfate-reducing bacterium closely related to *D. vulgaris* Hildenborough was isolated from heavy-metal impacted lake sediments located in northern Illinois (Lake DePue). Initial characterization by the ESPP AEC and Functional Genomics Core (FGC) revealed differences in genome content and structure between this strain (DePue) and strain Hildenborough, despite a very high level of 16S rRNA sequence similarity (>99%). Phenotypic analyses of this strain by the AEC revealed significant differences in minimum inhibitory concentrations for a variety of compounds when compared with strain Hildenborough. Strain DePue exhibited greater tolerance towards Cr(VI) and increased sensitivity to nitrate. Small differences were observed in growth rates, although not sensitivity, for sodium between the two strains. Genome sequencing of strain DePue by the DOE Joint Genome Institute indicated that the majority of genes (approximately 90%) share a high level of similarity (>98%) to genes found in strain Hildenborough. However, the genome of strain DePue exhibits multiple genome inversions and rearrangements, as well as the presence of a several hundred novel genes not found in strain Hildenborough. Current analyses by the ESPP Computational Core (CC) verified that strain DePue lacks at least six phage regions found in strain Hildenborough, but also suggests at least two unique phage regions, one of which contains putative multi-drug efflux genes.

Further curation of the genome by the ESPP CC, as well as mutant analysis by the FGC and AEC should inform the basis for increased metal-tolerance of strain DePue and metal-resistance among *Desulfovibrio* in general.

7 GTL

Web Tools for Revealing Relationships Among Strains, Taxa, and Communities

T.G. Lilburn,[1] S.H. Harrison,[2]* J.R. Cole,[2] P.R. Saxman,[3] and **G.M. Garrity**[2] (garrity@msu.edu)

[1]American Type Culture Collection, Manassas, Virginia; [2]Michigan State University, East Lansing, Michigan; and [3]University of Michigan, Ann Arbor, Michigan

Project Goals: The goals of this project are to develop and deploy tools that support analyses and visualizations of extremely large sequence data sets used in phylogenetic reconstructions. Current efforts are focused on validation of the self-organizing self-correcting classifier developed earlier and deployment of the tool as a web service that integrates with the RDP-II project.

Statistical approaches to understanding the species richness of prokaryotic communities in diverse environments indicate that there are thousands of yet to be cultured species (1, 4). Typically, putative members of these communities are classified and identified based on 16S rRNA genes in the extracted environmental DNA. These sequences are known as "environmental clones" in order to distinguish them from sequences from cultured organisms. The environmental clone sets are usually compared with publicly available sequences and projected as trees. In recent years the percentage of environmental clone 16S rRNA sequences in GenBank has increased from 67% to 80%. In our efforts to maintain the nomenclatural taxonomy, it has become clear that the preponderance of environmental clones is creating difficulties for researchers. We have applied our Taxomatic tool to resolving the phylogenetic taxonomy of the prokaryotes and to explore the effects of environmental clones on current classification and identification methods.

For illustrative purposes, we turn to Wagner's recently proposed "super phylum" (5) that encompasses three recognized phyla (the *Verrucomicrobia*, the *Planctomycetes*, and the *Chlamydiae*), the *Lentisphaerae* (currently part of the *Verrucomicrobia*) and two groups that contain no cultured representatives (the OP3 candidate phylum and the *Poribacteria*). When we searched GenBank for SSU rDNA sequences affiliated with the six groups and compared them with a comparable set of sequences obtained from the RDP-II database, the differences were startling. We retrieved 3,568 SSU rDNA sequences from GenBank and 4,595 from the RDP-II database. The intersection of these two sets included only 2,160 sequences; 2,435 sequences were identified as members of this group only by the RDP-II and 1,408 were identified only by GenBank (Figure 1). Clearly, the interpretation of a community analysis would differ depending on the data set used to classify the sequences obtained from the environment.

* Presenting author

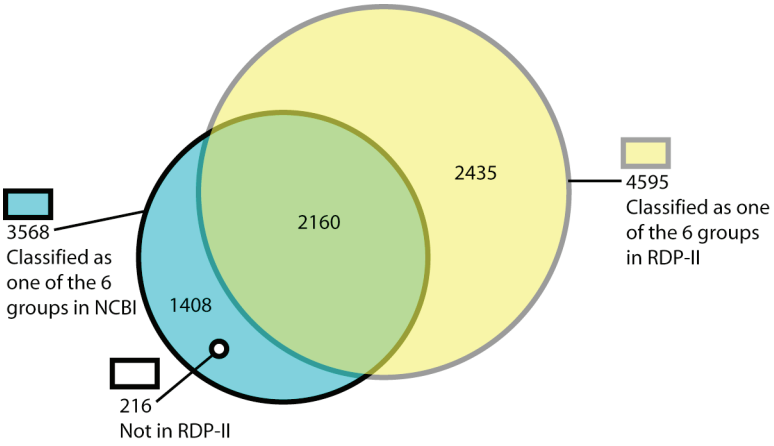


Figure 1: Showing the distribution of sequences identified as members of the six groups proposed to be members of a super phylum. Overall, 6,003 sequences were retrieved from the two databases, but only 2,160 sequences were identified by both databases as members of the six groups.

A comparison of the two classifications could help explain why there is such a marked difference in the data sets. The comparison was done using the Taxomatic, a web service built around the SOSCC algorithm (3). The algorithm produces visualizations of classifications as heat maps in R. Because biologists may not be familiar with the R or S statistical programming environments in which the algorithm runs, we have developed a web service to produce heat maps on demand. The web service has three components. The first allows users to assemble a data set and/or select a starting classification using an already-familiar interface – the RDP-II web site. The second consists of the compute engine – in essence the SOSCC algorithm running on a separate server. The third presents the results in the form of a dynamic heat map that allows users to drill down through the hierarchy and to link to external resources using digital object identifiers through NamesforLife semantic resolution services. Messaging and operations are handled by web service definition language documents in combination with the simple object access protocol. This architecture provides a robust and scalable interface to large dynamic information resources, such as the RDP-II.

Performance-tuning of the SOSCC feature set is centered around usage cases, including benchmarking, subsampling, comparing alternative hierarchies, and charting novel sequences. In the simplest case, users can upload a list of sequence identifiers to produce a publication-ready heatmap of their data. More advanced options support input of predefined or user-derived classifications. Output of intermediate steps and documentation in the SOSCC classification process are also available for user inspection. Scalability issues are also being addressed so that SOSCC web services will function as the dataset grows.

Our results show that user classifications captured by GenBank are distorting our picture of microbial diversity. Many sequences in GenBank have been placed into the taxonomic hierarchy by the submitting authors at the time of deposit, based on a BLAST-nearest neighbor approach. We have previously reported that BLAST cannot reliably determine the nearest rRNA sequence, most likely because of the high degree of similarity among all rRNA sequences (2). Moreover, since researchers apply the classification (and other annotations) from the BLAST nearest neighbor to their sequence(s), and since it is likely that the nearest neighbor is an environmental clone, each iteration of this process moves the classification another step away from any data anchored to a cultivated organism. Annotation transfer relies completely on the accuracy of prior work and therefore can

lead to misclassifications or misidentifications. This strategy also has the effect of propagating and amplifying prior errors, especially in the case of taxa for which there are few cultured representatives. In contrast, the RDP-II has implemented an on the fly Bayesian classifier to place sequences into the classification during downloading from GenBank. The classifier is trained with a carefully validated set of sequences from the nomenclatural taxonomy, so the RDP-II (and thus the Taxomatic web service) is able to provide users with more reliable and up-to-date assessments of community membership and more accurate identifications of new taxa, based on the SSU rDNA sequences.

References

1. **Acinas, S. G., V. Klepac-Ceraj, D. E. Hunt, C. Pharino, I. Ceraj, D. L. Distel, and M. F. Polz.** 2004. Fine-scale phylogenetic architecture of a complex bacterial community. Nature **430:**551-4.
2. **Cole, J. R., B. Chai, R. J. Farris, Q. Wang, S. A. Kulam, D. M. McGarrell, G. M. Garrity, and J. M. Tiedje.** 2005. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res **33:**D294-6.
3. **Garrity, G. M., and T. G. Lilburn.** 2005. Self-organizing and self-correcting classifications of biological data. Bioinformatics **21:**2309-2314.
4. **Schloss, P. D., and J. Handelsman.** 2006. Toward a census of bacteria in soil. PLoS Comput Biol **2:** e92.
5. **Wagner, M., and M. Horn.** 2006. The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. Curr Opin Biotechnol **17:**241-9.

8 GTL

High Quality Microbial Finishing at JGI

Alla Lapidus[1]* (alapidus@lbl.gov), Eugene Goltsman,[1] Steve Lowry,[1] Hui Sun,[1] Alicia Clum,[1] Stephan Trong,[1] Pat Kale,[1] Alex Copeland,[1] Patrick Chain,[2] Cliff Han,[3] Tom Brettin,[3] Jeremy Schmutz,[4] and Paul Richardson[1]

[1]DOE Joint Genome Institute (JGI) Production Genomics Facility, Walnut Creek, California; [2]JGI-Lawrence Livermore National Laboratory, Livermore, California; [3]JGI-Los Alamos National Laboratory, Los Alamos, New Mexico; and [4]JGI-Stanford, Stanford, California

Project Goals: To provide JGI collaborators with the highest possible quality complete microbial genomes.

The value of complete microbial genome sequence is established and appreciated by scientific community. A finished genome represents the genome assembly of high accuracy and quality (with no gaps), verified and confirmed through a number of computer and lab experiments. Several yeas ago JGI has established a set of high standards for the final microbial assembly and has been strictly following them thereafter.

More than 100 microbial projects have been completed since that time within the framework of the JGI's portfolio (DOE GTL program, DOE Microbial program and the Community Sequencing Program). Progress in DNA sequencing technology, design of new vectors for library construction, improvements in finishing strategies and tools, as well as the availability of a number of assemblers and advanced methods for OFR finding and genome annotation have significantly reduced the time required for genome closure. Despite this fact, complexity and speed of genome closure depends on the quality of DNA received, the whole genome shotgun libraries produced from this DNA, GC content of the genome, the size and frequency of identical or nearly identical repetitive structures,

* Presenting author

and the amount of regions that can not be cloned or had to clone in *E. coli*. The whole genome finishing/assembly improvement pipeline will be presented showing the lab approaches and computational finishing techniques developed and implemented at JGI for finishing the large number of microbial projects in the queue. We also will present our progress in completing metagenomic projects. A number of projects for which the combination of different sequencing technologies (Sanger and 454) and finishing strategies were used will also be presented.

This work was performed under the auspices of the U.S. Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under Contract No. W-7405-ENG-36.

9 GTL

Evolution of Energy Metabolism in the *Geobacteraceae*

J.E. Butler* (jbutler@microbio.umass.edu), N.D. Young, D. Kulp, and **D.R. Lovley**

Department of Microbiology, University of Massachusetts, Amherst, Massachusetts

Project Goals: To determine the evolution of energy metabolism in the *Geobacter* family.

To better understand the pathways of energy metabolism in the *Geobacteraceae* family of Fe(III) reducers and electricigens, gene conservation, gene loss, and horizontal gene transfer were determined for the whole genomes of six species in the family. According to 16S rDNA phylogeny, these six species fall into two clades – the freshwater *Geobacter* clade that includes the *Geobacter* genus and *P. propionicus*, and the marine *Desulfuromonas* clade that includes *D. acetoxidans* and *P. carbinolicus*. We sought to determine the gene set shared by all members of this family, as well as to explain the presence of *Pelobacter* species, both primarily fermentative organisms, in both clades of the *Geobacteraceae*. The set of 529 proteins that were found in a single copy in each of the genomes were concatenated and used to model the phylogeny of the family. This super-tree confirmed that there are two distinct clades of *Geobacteraceae* species, and that there are *Pelobacter* species in each clade. Analysis of the genes conserved in all six genomes showed there were 713 families common to all six species. A complete TCA cycle was present in each species, although the *Pelobacters* had non-homologous isocitrate dehydrogenase and fumarase enzymes. In addition, the enzyme complexes of the inner membrane electron transport chain were generally well conserved. All species contained at least one NADH dehydrogenase, a succinate dehydrogenase, and at least one ATP synthase. Notably, only the *Pelobacters* lacked a cytochrome *bc* complex, which is predicted to move electrons from the inner membrane out to the periplasm. Surprisingly, although each genome contains at least 40 *c*-type cytochrome genes, only one of these genes was found to be conserved in all 6 members of this family, the catalytic subunit of the nitrite reductase. None of the cytochromes previously shown to be required for Fe(III) reduction or electricity generation in *G. sulfurreducens* was conserved in all the genomes. Analysis of the *Pelobacter* species indicated that they have lost a similar set of genes when compared to the non-fermenting species of the family, including: acetate transporters, all of the hydrogen-oxidizing hydrogenases, the formate dehydrogenases, and most of the *c*-type cytochromes. Analysis of gene gain in the *Pelobacter* species indicate that both gained a small cluster of dehydrogenases that allow them to use butanediol and acetoin. However, they metabolize these substrates with 2 different reaction pathways, one involving carbon-fixation and the inner membrane electron transport complexes, and the other dependent on only cytosolic proteins. These results indicate that the proto-*Geobacter* was

likely a respiring species, dependent on oxidation of carbon compounds coupled to a typical chain of electron transport complexes. The ability to ferment arose two separate times in this family.

10

Establishing Potential Chloroplast Function Through Phylogenomics

Sabeeha Merchant[1]* (merchant@chem.ucla.edu), Steven Karpowicz,[1] Arthur Grossman,[2] Simon Prochnik,[3] and Dan Rokhsar[3]

[1]Department of Chemistry and Biochemistry, University of California, Los Angeles, California; [2]Department of Plant Biology, The Carnegie Institution, Stanford, California; and [3]DOE Joint Genome Institute, Walnut Creek, California and Center for Integrative Genomics, Department of Molecular and Cell Biology, University of California, Berkeley, California

Project Goals: The structure and metabolism of chloroplasts from green algae and land plants is strikingly well conserved, especially with respect to components of the photosynthetic apparatus and the factors involved in its assembly and maintenance. Analyses of genome sequences from various algae and plants have suggested that novel components of the chloroplast might be revealed by comparative phylogenomics. *Chlamydomonas* has served historically as a powerful model organism for the discovery of photosynthetic components and fundamental aspects of chloroplast genome expression. Therefore, we sought to use the complete set of *Chlamydomonas* protein annotations as a central data set for divining highly conserved components in the chloroplast. A list of 189 proteins conserved in organisms with green chloroplasts was generated. The goal of the project is to assess the functions of proteins on this list whose functions are not yet known. This would involve phenotypic analysis of mutants, sub-cellular locations of proteins, identification of interaction partners, and patterns of expression.

The structure and metabolism of chloroplasts from green algae and land plants is strikingly well conserved, especially with respect to components of the photosynthetic apparatus and the factors involved in its assembly and maintenance. Analyses of genome sequences from various algae and plants have suggested that novel components of the chloroplast might be revealed by comparative phylogenomics. *Chlamydomonas* has served historically as a powerful model organism for the discovery of photosynthetic components and fundamental aspects of chloroplast genome expression. Therefore, we sought to use the complete set of *Chlamydomonas* protein annotations as a central data set for divining highly conserved components in the chloroplast. The putative orthologs of *Chlamydomonas* genes from organisms with full genome sequences (*Arabidopsis, Physcomitrella,* diatoms, *Ostreococcus spp*., cyanobacteria, nematode, slime mold, human, *Neurospora*, *Phytophthora*, archaea, and non-photosynthetic bacteria) were determined based on a mutual best hits approach with WU-blast (Version 2). Next, we attempted to add only close paralogs or "inparalogs" (genes that have duplicated since speciation [1]) to pairs of orthologs, while excluding 'outparalogs', although this is hard at such large evolutionary distances. The combination of paralogs and orthologs generated clusters of proteins that were presumably represented in the ancestor by a single gene. Having made clusters of genes from organisms across the kingdoms of life, we were able to ask for clusters that contained proteins from certain combinations of organisms, enabling U.S. to generate a series of lists of proteins conserved in different green organisms but not present in non-photosynthetic eukaryotes or prokaryotes from the protein clusters. We started with a list of 914 proteins conserved between *Arabidopsis* and *Chlamydomonas* but not present in the non-photosynthetic organisms used in the study, and then restricted the lists sequentially by inclusion of additional photosynthetic organisms, in order: *Physcomitrella*,

* Presenting author

Ostreococcus, diatoms, *Cyanidioschizon merolae* and cyanobacteria. For *Ostreococcus*, diatoms and cyanobacteria, where genome sequences of more than one species are available, parallel lists requiring the presence of orthologs in "at least one" or "both" species were generated. A working list consisting of 189 *Chlamydomonas* proteins in 145 clusters, consisting of proteins conserved in *Arabidopsis*, *Chlamydomonas*, *Physcomitrella*, *Ostreococcus* and at least one diatom was analyzed in detail for a) known or predicted protein functions, b) predicted or experimentally-determined protein localization and c) pattern of expression of the Arabidopsis homolog. Most of the proteins on the list are indeed plastid localized or predicted to be so, and in most cases the pattern of expression is compatible with a function in photosynthesis or other anabolic pathways. There are no false positives on this list among the known proteins (~ 50% of the total), which include proteins unique to photosynthesis or the biogenesis of the photosynthetic apparatus, such as phosphoribulokinase, OEE proteins, Rubisco methyl transferase, HCF164, and enzymes involved in tetrapyrrole metabolism such as DVR, CHLD, GUN4, HMOX. Components unique to chloroplast metabolism were also identified among the known proteins, including proteins in the Vitamin E and Vitamin C biosynthesis pathways, and dihydropicolinate reductase in a plant specific lysine biosynthesis pathway. Plastid-specific isozymes of the pyruvate dehydrogenase complex were also selected. On this basis we conclude that the unknown components are likely to represent chloroplast-localized proteins with functions in photosynthesis or other key chloroplast metabolic pathways. Furthermore, motif analysis suggests that some of some unknown proteins may function in redox regulation, metabolite exchange, or genome maintenance/expression. Our conservative phylogenomics strategy is likely to have identified many novel proteins involved in photosynthesis, with few false hits. While the computational analysis was facilitated by the whole genome sequence of *Chlamydomonas*; the experimental accessibility of the organism means these predictions can be tested very readily.

Research sponsored by U.S. DOE Energy Biosciences, USDA NRI Plant Biochemistry, and NSF Plant Genome.

Reference

1. Sonnhammer, E.L. and E.V. Koonin, Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet., 2002. 18(12): p. 619-20.

11

Beneficial Effects of Endophytic Bacteria on Biomass Production by Poplar

Safiyh Taghavi and Daniel van der Lelie* (vdlelied@bnl.gov)

Biology Department, Brookhaven National Laboratory, Upton, New York

Project Goals: The aim of this project is to understand the beneficial interaction between poplar and its endophytic bacteria. The association of endophytic bacteria with their plant hosts have been shown to have a growth-promoting effect for many different plant species. Endophytic bacteria have several mechanisms by which they can promote plant growth and health. These include the production of phytohormones or enzymes involved in growth regulator metabolism such as ethylene, 1-aminocyclopropane-1-carboxylic acid (ACC) deaminase, auxins, indoleacetic acid (IAA) or cytokinins. These mechanisms are of direct importance to the DOE mission of carbon sequestration through biomass production. In addition, endophytic bacteria can help their host plants to overcome the phytotoxic effects caused by environmental contamination, which is of direct relevance for waste management and pollution control via phytoremediation technologies.

Introduction

The association of endophytic bacteria with their plant hosts have been shown to have a growth-promoting effect for many different plant species. Endophytic bacteria have several mechanisms by which they can promote plant growth and health. These include the production of phytohormones or enzymes involved in growth regulator metabolism such as ethylene, 1-aminocyclopropane-1-carboxylic acid (ACC) deaminase, auxins, indoleacetic acid (IAA) or cytokinins. These mechanisms are of direct importance to the DOE mission of carbon sequestration through biomass production. In addition, endophytic bacteria can help their host plants to overcome the phytotoxic effects caused by environmental contamination, which is of direct relevance for waste management and pollution control via phytoremediation technologies.

Scientific approach

Recent research by our group has illustrated the potential of endophytic bacteria to increase the net primary biomass production of the host *Populus* tree. The goal of our research is to identify specific strains of endophytic bacteria that improve the growth and carbon sequestration potential of *Populus*. We therefore conducted a high-throughput screen of approximately 100 bacterial endophyte strains to identify those strains with the greatest impact on *Populus* net primary productivity. As an example, the results of inoculation of poplar with 8 different endophytic strains are presented. The most significant stimulation in growth was observed with *Enterobacter* sp. 638*, followed by *Burkholderia cepacia* L.S.2.4 and *Stenotrophomonas maltophilia* R551-3*. Some strains had no significant effect on biomass production as compared to non inoculated control plants. This was the case for *Enterobacter* sp. R558-1, *Pseudomonas putida* W619*, *Serratia proteamaculans* 568* and plants inoculated with the soil bacterium *Ralstonia metallidurans* CH34 (control). Interestingly, plants inoculated with *Methylobacterium populi* BJ001* showed a strong reduction in growth, despite the fact that this strain was isolated as an endophyte from poplar tissue cultures.

To better understand the interactions between poplar and its endophytic bacteria we initiated in collaboration with DOE's JGI the full genome sequencing of 5 endophytic strains (marked by *). A first analysis of the draft genome sequences resulted in the identification of several functions that would allow the endophytic bacteria to interact with the development of their poplar host. Several strains contained a copy of a 1-aminocyclopropane-1-carboxylic acid (ACC) deaminase gene, seemed to be able to produce indoleacetic acid (IAA), or metabolize phenyl acetic acid. In addition, *Pseudomonas putida* W619 was shown to contain the uptake carrier for 4-amino butyrate, another important plant hormone. Genome comparison between the endophytes and closely related non-endophytic strains from the same species should provide U.S. with valuable insights about the essential functions for successful endophytic colonization by these bacteria of their poplar host.

* Presenting author

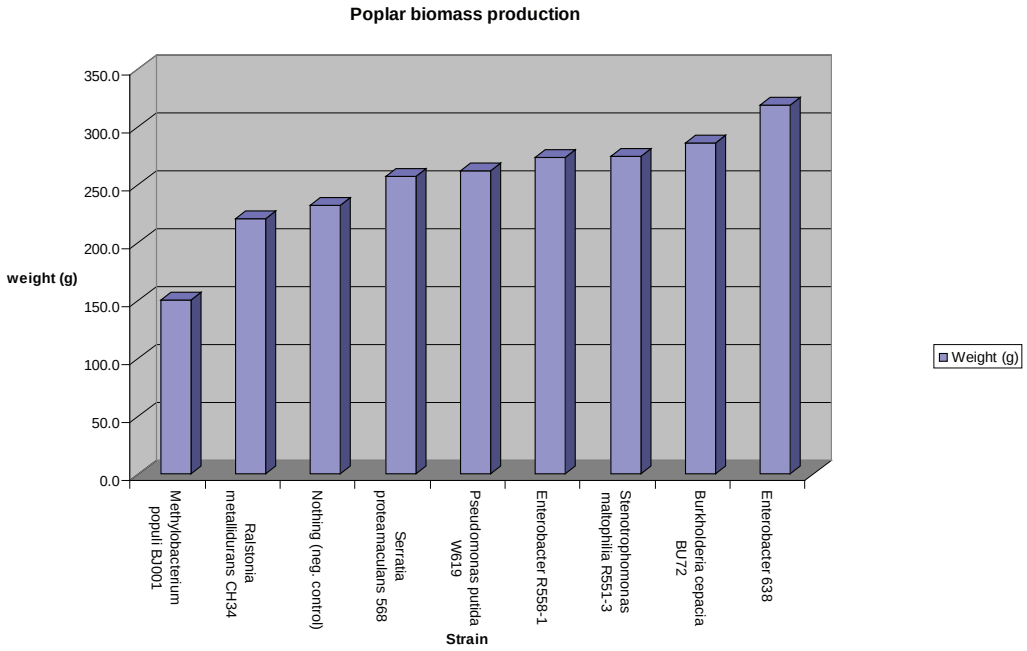


Figure 1: beneficial effects of endophytic colonization on the biomass production of their poplar host. Results are the average of 7 poplar trees per condition.

Future work

We will test the hypothesis that the interacting genomes of bacterial endophytes and *Populus* not only increase *Populus* biomass production, but also increase partitioning of resources into long-lived, i.e. recalcitrant forms of carbon. In parallel, endophyte properties that are hypothesized to be important for colonization and plant growth promotion will be investigated, including capacity for production of extracellular hydrolytic enzymes, nitrogen-fixing enzymes, and low molecular weight compounds such as phytohormones. Also, the dynamics of endophytic colonization will be explored using GFP-expressing endophytic strains. With these data in hand we will embark upon a systems biology approach to better understand the interaction between endophytic bacteria and their *Populus* host. The resulting comprehensive view of the endophyte-*Populus* interacting genomes has the potential to be used in developing recommendation in use of endophyte inoculant to increase carbon sequestration in *Populus* plantations.

Acknowledgement

This work was supported by the U.S. Department of Energy, Office of Science, BER, project number KP1102010. This work was also funded under Laboratory Directed Research and Development project number LDRD05-063. Sequencing of the endophytic genomes is been carried out at the Joint Genome Institute (JGI) under the auspices of the U.S. Department of Energy's Office of Science, Biological and Environmental Research Program.

Section 2

Microbial Community Sequencing and Analysis

12 GTL

Structure and Dynamics of Natural Low-Diversity Microbial Communities

Jillian F. Banfield[1,5]* (jill@eps.berkeley.edu), Vincent Denef,[1] Nathan VerBerkmoes,[2] Paul Wilmes,[1] Gene Tyson,[1] John Eppley,[1] Genevieve DiBartolo,[1] Daniela Goltsman,[1] Anders Andersson,[1] Chris Belnap,[1] Brett J. Baker,[1] Linda Kalnejais,[1] A. Pepper Yelton,[1] D. Kirk Nordstrom,[3] Eric E. Allen,[1] Rachel Whitaker,[1] Sheri Simmons,[1] Manesh Shah,[2] Michael Thelen,[4] Gary Andersen,[5] and Robert Hettich[2]

[1]University of California, Berkeley, California; [2]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [3]U.S. Geological Survey, Boulder, Colorado; [4]Lawrence Livermore National Laboratory, Livermore, California; and [5]Lawrence Berkeley National Laboratory, Berkeley, California

Project Goals: The interdisciplinary research in our GTL project is motivated by the desire to understand how microbial communities assemble, the interplay amongst community members, and the mechanisms of microbial adaptation and evolution. Chemoautotrophic acid mine drainage (AMD) microbial biofilms have proven particularly tractable for these analyses due to their comparatively low species richness. The goal of this subproject is to develop experimental and computational approaches for the comprehensive characterization of the proteome of the AMD system to investigate the nature of the gene expression and conservation amongst the various microbial members of this consortium. Proteomic information will be integrated with genomic and biochemical datasets to help elucidate the structure and activity of microbial communities in their natural environmental context.

The interdisciplinary research in our GTL project is motivated by the desire to understand how microbial communities assemble, the interplay amongst community members, and the mechanisms of microbial adaptation and evolution. Our approach is to use cultivation-independent methods to study the structure and activity of microorganisms in their natural environmental context. Chemoautotrophic acid mine drainage (AMD) microbial biofilms have proven particularly tractable for these analyses due to their comparatively low species richness. Recent advances have relied upon the development and application of community genomic and proteomic methods, as well as new contextual information provided by geochemical analyses and ultrastructural characterization.

Genomic studies reveal that AMD biofilms are typically dominated by near clonal bacterial populations. Spatial and temporal analyses indicate essentially a single species of *Leptospirillum* group III, the only organism capable of nitrogen fixation, across a diversity of environment types. In contrast, different biofilms are dominated by one of a set of *Leptospirillum* group II genomes formed by homologous recombination between two end member genome types. Analyses of biofilm growth stages suggest selection for a UBA *Leptospirillum* group II type during initial colonization, introduction of archaea and fungi in intermediate succession stages, and dominance by the 5-way CG *Leptospirillum* group II genome type in mature biofilms.

* Presenting author

Within bacterial populations, the predominant form of genomic heterogeneity is in gene content. These differences are particularly prevalent in regions impacted by phage and plasmid integration. In some cases, these regions encode key functions, including cytochrome variants and quorum sensing genes. Our findings further indicate that subsets of these genes are under selection. Variation in metabolic potential resulting from gain or loss of phage-related genes is inferred to be particularly important in diversification of otherwise near clonal *Leptospirillum* group III populations. Predominance of a single genome type for each bacterial species may reflect strong clonal expansion events, particularly early in the early colonization of the air-solution interface.

Archaea are numerically less abundant in biofilm communities. Archaeal populations typically have complex genome pools consisting of combinatorial variants, and also exhibit heterogeneity in gene content concentrated in phage insertion regions. The high degree of sequence-level variation is maintained by rapid rates of homologous recombination, possibly ensuring a continuum of adaptation potential. This may be important, given that they appear in successional stages where microenvironmental heterogeneity is likely the result of established biofilm architecture.

In addition to assembly of microbial consortia in response to physical and chemical conditions and biofilm growth stage, recent evidence suggest that viral predation is an important selective force that shapes microbial consortia and drives their evolution. We find evidence for dramatic crashes in biofilm communities, possibly caused by phage blooms, accompanied by a switch in the dominant bacterial strain type. Periodic decimation of the dominant populations is not unexpected (*kill the king*), given ongoing microbial evolution to outwit phage counterbalanced by phage evolution to evade host defenses.

Microbial genomes may provide insight into the mechanisms of phage defense and record information about the recent history of phage predation. All of the microorganisms in the AMD biofilms contain at least one region of short tandem repeats separated by similar length spacer sequences (CRISPR), accompanied by a set of CRISPR-associated proteins; previous studies have suggested that this comprises a microbial immune system. We find very high levels of population heterogeneity in the spacers between tandem repeats, consistent with rapid diversification of the inventory of possible RNAi-like molecules available to silence foreign DNA. A significant subset of the spacers shows sequence similarity to transposase genes, prophage genes, and unassembled sequencing reads (possible derived from phage). In combination with other data, results support a role of spacers in phage defense. Extremely rapid CRISPR evolution is expected if the region is responding to a rapidly changing selection pressure associated with phage predation. Bacterial CRISPR-associated proteins are some of the most abundant proteins in the biofilms, reinforcing the importance of these large genomic loci to organism survival. Ongoing parallel studies of phage communities, in combination with genomic and proteomic analysis of bacteria and archaea, will be vital for development of a more detailed understanding of microbial community dynamics.

This research sponsored by the U.S. DOE-BER, Genomics:Genomics:GTL Program.

13 GTL

A Novel Binning Approach and Its Application to a Metagenome From a Multiple Extreme Environment

N. Maltsev[1]* (maltsev@mcs.anl.gov), M. Syed,[1] A. Rodriguez,[1] B. Gopalan,[2] and **F. Brockman**[2]

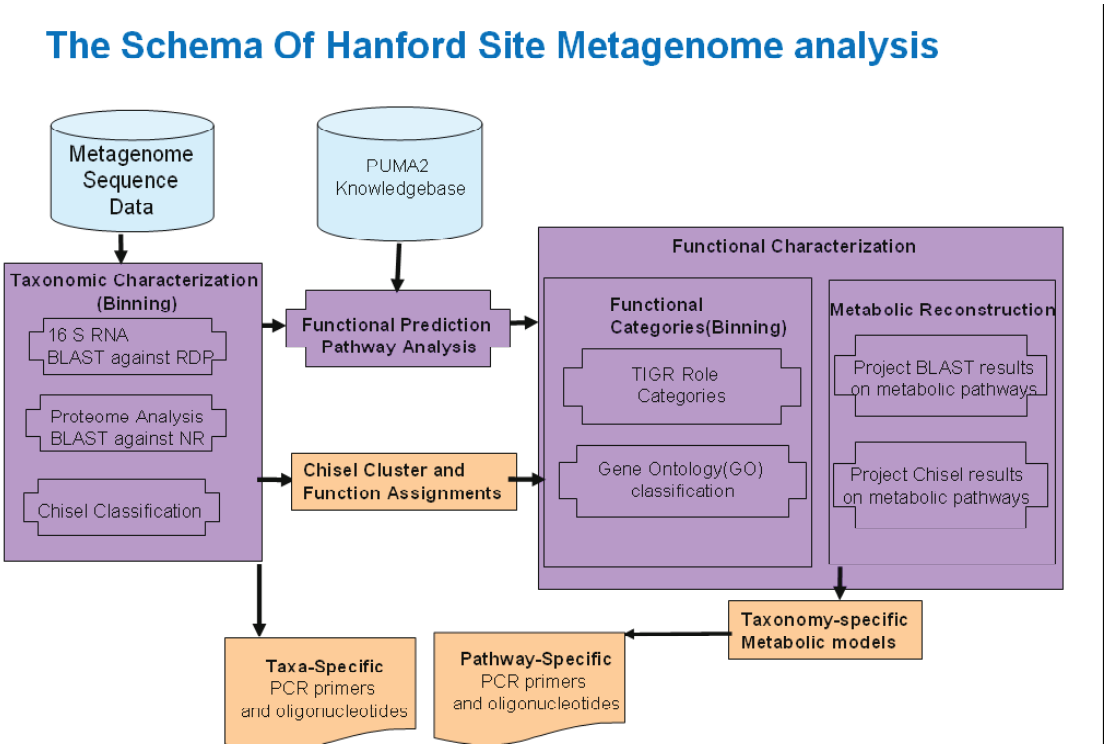
[1]Argonne National Laboratory, Argonne, Illinois and [2]Pacific Northwest National Laboratory, Richland, Washington

Project Goals: A metagenome of the living microbial community present in low biomass subsurface sediments beneath a leaking high-level radioactive waste tank at the DOE Hanford Site was analyzed using advanced bioinformatics methods. Besides very high levels of radiation this microbial community was also subjected to high temperatures and dessication, extremely high concentrations of chromium and nitrate, and alkaline conditions. The great majority of microbes in most natural environments will be represented in metagenome sequence by a limited number of genes which do not assemble, due to substantial community diversity and limited sequencing depth. Therefore, a critical need is more accurate and informative binning of genes into taxonomic groups, to enable improved reconstruction of the metabolic and physiological processes operating in the community. We have developed a new approach for binning metagenome sequences, and appled it to a low-biomass microbial community exposed for several decades to multiple extreme conditions. We have discovered that many of the proteins found in this metagenome show homology to those found in extremophilic microbes, indicating that the community has undergone systems-level changes.

A metagenome of the living microbial community present in low biomass subsurface sediments beneath a leaking high-level radioactive waste tank at the DOE Hanford Site was analyzed using advanced bioinformatics methods. Besides very high levels of radiation this microbial community was also subjected to high temperatures and dessication, extremely high concentrations of chromium and nitrate, and alkaline conditions. The great majority of microbes in most natural environments will be represented in metagenome sequence by a limited number of genes which do not assemble, due to substantial community diversity and limited sequencing depth. Therefore, a critical need is more accurate and informative binning of genes into taxonomic groups, to enable improved reconstruction of the metabolic and physiological processes operating in the community. We have developed a new approach for binning metagenome sequences, and appled it to a low-biomass microbial community exposed for several decades to multiple extreme conditions. We have discovered that many of the proteins found in this metagenome show homology to those found in extremophilic microbes, indicating that the community has undergone systems-level changes.

Hanford site metagenome. Biomass in the sediments is present in very low quantities (~10,000 cells per gram), about one-millionth the levels routinely found in soils. In order to exclude the large fraction of dead cells and to provide adequate biomass for library construction, many culture enrichments were pooled. Low-coverage shotgun sequencing was performed by the DOE Production Genomics Facility. Metagenome sequence obtained from pooled enrichments from the extreme contamination and low contamination zones were separately analyzed using bioinformatics approaches to achieve the following: 1. Determine the taxonomy of organisms selected for by the extremophilic conditions 2. Reconstruct major physiological properties of the microbial community from available genomic data and 3. Identify the sequence features associated with extremophilic phenotypes.

* Presenting author



1. Taxonomic Profiling (Binning) was done using**:** a) Phylogenetic analysis of the 16S rDNA b) Taxonomic analysis of the BlastX results and c) Identification of taxonomic variations of enzymes using the Chisel system being developed by our group. This unique step allows increasing resolution and reliability of predictions of enzymatic capabilities characteristic for particular taxonomic groups in the samples.

2. Reconstruction of Physiological Profiles for taxonomic groups of the microbial community. The gene function predictions were based on the results of analysis of metagenomes using Chisel [http://compbio.mcs.anl.gov/CHISEL] for identification of taxonomy and phenotype-specific variations of enzymatic sequences, as well as by traditional bioinformatics tools (e.g. BlastX, InterPro, Blocks). The predicted functions attributed to particular organisms or taxonomic groups were projected onto the library of metabolic pathways from the EMP and KEGG databases. Reconstructions of signal transduction and transmembrane transport systems were also performed.

Results

Sixty-three 16S rDNA and 13,388 protein sequences from the extreme and low contamination zones were analyzed. According to the analysis of the 16S rDNA the extreme contaminatin zone was dominated by *Deinococcus-Thermus, Actinobacteria* and *Firmicutes*. The *Actinobacteria* were represented by *Micrococcineae, Propionibacterineae*, *Corynebacterineae* and *Frankineae* suborders. Taxonomic analysis of Chisel results and BlastX hits has confirmed that this population was dominated by the *Deinococcus-Thermus* phylum (~60% of ORFs) and *Actinobacteria* (~37% of ORFs). A significant number of homologs to sequences from extremophilic organisms were identified. The low contamination zone was dominated by *Proteobacteria* and *Actinobacteria*; 6 of 9 16S rDNA sequences were attributed to *Actinobacteria*. Analysis of protein sequences using Blast and Chisel has attributed ~ 55% of translated ORFs to *Proteobacteria* phylum and 40% to *Actinobacteria.*

Metabolic Reconstructions from sequence data were done using the gene function predictions based on Chisel results and conventional bioinformatics tools. Due to low coverage of genomes in the samples, in the majority of cases reconstruction of physiological profiles for individual species was impossible. Therefore hierarchical reconstructions for higher taxonomic groups (e.g. genus, order) were performed. The predicted enzymes in the most abundant groups (*Actinomycetales* and *Bacillus*) corresponded to the core metabolic pathways. An interesting finding was the identification of enzymes in the streptomycin biosynthetic pathway in *Actinobacteria.* Streptomycin is known to induce streptomycin-dependent error-prone protein biosynthesis, that may be advantageous for microorganisms residing in extreme conditions.

Evolutionary Analysis of genomes of extremophilic organisms. The Chisel analysis of the extreme contamination zone metagenome identified 543 enzymatic sequences corresponding to 263 distinct enzymatic functions. The predominant taxonomic groups of organisms identified in the sample were *Actinomycetales* (28%, corresponding to 152 Chisel predictions) and *Bacillus* (22%, corresponding to 122 Chisel predictions). Other predicted groups included a number of hits from extremophilic organisms: *Deinococcus*, *Euryarchaeota,* and *Symbiobacterium thermophilum*. These results match the results predicted by 16S rDNA analysis of this data. Chisel allows for further investigation of this metagenome by supporting the design of taxonomy-specific oligonucleotides for messenger RNA-targeted fluorescence in situ hybridization (FISH) studies. These degenerative oligonucleotides are based on the alignments of sequences corresponding to taxonomy-specific Chisel clusters.

The taxonomic profile of the microbial community identified in the extreme contamination zone shows a surprising similarity to the community in untreated and low radiation (0.5 MRad) treated soils of the Atacama desert (Rainey et al., 2004). Both populations were dominated by *Actinobacteria, Deinococcus* and *Firmicutes.* This observation leads to the suggestion that microbial populations residing in extremophilic natural environments are pre-conditioned for adapting to and surviving new, human-caused extremophilic conditions. To test this hypothesis we further analyzed the taxonomy-specific variations of enzymes identified by Chisel in our metagenomes, using high-resolution bioinformatics tools developed by our group (e.g. Dragonfly and Phyloblocks) for evolutionary analysis of protein sequences. Our analysis shows divergent evolution of enzymatic functions that lead to the emergence of Actinobacterial and Deinococcal variations of some essential enzymes of glycolysis, nucleotide biosynthesis, DNA repair systems, and others. Many of these enzymes also show subsequent convergent evolutionary changes characteristic for extremophilic microbes from different taxonomic groups. We conclude that in the course of adaptation to the conditions in the Hanford sediments, the community has undergone systems-level changes spanning multiple biological functions.

* Presenting author

14 GTL

Insights into Stress Ecology and Evolution of Microbial Communities from Uranium-Contaminated Groundwater Revealed by Metagenomics Analyses

Christopher L. Hemme,[1,6,8]* Ye Deng,[1] Terry Gentry,[6] Liyou Wu,[1] Matthew W. Fields,[2,8] David Bruce,[3] Chris Detter,[3] Kerrie Barry,[3] David Watson,[6] Paul Richardson,[3] James Bristow,[3] Terry C. Hazen,[4,8] James Tiedje,[5] Eddy Rubin,[3] **Adam P. Arkin**[7,8] (aparkin@lbl.gov), and Jizhong Zhou[1,8]

[1]Institute for Environmental Genomics, Department of Botany and Microbiology, University of Oklahoma, Norman, Oklahoma; [2]Department of Microbiology, Miami University, Oxford, Ohio; [3]DOE Joint Genome Institute, Walnut Creek, California; [4]Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, California; [5]Center for Microbial Ecology, Michigan State University, East Lansing, Michigan; [6]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [7]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; and [8]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

One of the central goals of ESPP is to understand the responses of microbial communities to various stresses within the context of field applications. Towards this goal, we are sequencing groundwater microbial communities with manageable diversity and complexity (~10-400 phylotypes) at the U.S Department of Energy's Environmental Remediation Science Program (ERSP)-Field Research Center (FRC), Oak Ridge, TN. The microbial community has been sequenced from a groundwater sample (FW106) contaminated with very high levels of nitrate, uranium and other heavy metals and pH ~3.7. Consistent with trends expected in stressed ecosystems, the metagenome reveals a community of low species and strain diversity dominated by a single *Frateuria*-like γ-proteobacteria with other γ- and β-proteobacteria present at low proportions. Metabolic reconstruction reveals specific adaptations to the geochemical conditions of FW106 including genes encoding metal resistance (*czcABC*, *czcD*, *cadA*, *merA*, *arsB*), denitrification, and solvent resistance (1,2-dichloroethene, acetone, butanol). In addition to the presence of these specific genes, certain resistance genes also appear to be overrepresented in the metagenome including genes from nitrate/nitrite transport (*narK*) and metal translocation (*czcABC*, *czcD*, *cadA*), likely due to a combination of gene duplication and lateral gene transfer. A screen for positive selection shows most of these genes to be under strong negative selection, suggesting that in the short term at least, the overabundance of these transporters provide a positive fitness benefit to the cell by increasing the rate of ion transport. SNP analysis revealed a low level of polymorphism with the overwhelming majority of SNP representing unique changes within the assembled reads, suggesting that the strains in the sample are largely clonal. A model is presented for the evolution of microbial communities under high-stress conditions. To understand the metabolic diversity of the groundwater microbial community, the microbial community (~400 phylotypes) from the background well at the FRC is also currently under sequencing.

* Presenting author

15 GTL

Changes in Microbial Community Structure During Biostimulation for Uranium Reduction at Different Levels of Resolution

C. Hwang,[1,8]* W.-M.Wu,[2] T.J. Gentry,[3] J. Carley,[4] S.L. Carroll,[4] D. Watson,[4] P.M. Jardine,[4] J. Zhou,[5,8] T.C. Hazen,[6,8] E.L. Brodie,[6,8] Y.M. Piceno,[6] G.L. Andersen,[6] E.X. Perez,[7] A. Masol,[7] C.S. Criddle,[2] and M.W. Fields[1,8]

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

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

Former radionuclide waste ponds at the ERSP-Field Research Center in Oak Ridge, TN pose several challenges for uranium bioremediation. The site is marked by acidic conditions, high concentrations of nitrate, chlorinated solvents, and heavy metals. Bacterial sulfate reduction can be an important process for the bio-reduction of *in situ* heavy metals, but little is known how potential stressors can impact bio-reduction activities at the cellular, population, and community levels. A goal of VIMSS is to characterize ecosystem responses at coordinated levels of resolution in order to predict cellular responses at DOE waste sites. Through VIMSS efforts, population and community level responses can be correlated with cellular responses from individual stress experiments, and this work allows for a more complete understanding of the system. The current work uses a series of re-circulating wells serve to create a subsurface bioreactor to stimulate microbial growth for *in situ* U(VI) immobilization (Wu et al. ES&T 40:3986-3995). Well FW-104 is the injection well for the electron donor (i.e., ethanol); well FW-026 is the extraction well for the recirculation loop; well FW-101 and FW-102 are the inner zones of biostimulation; and FW-024 and FW-103 are upstream and downstream wells, respectively, which are the outer protective zones. Microbial community composition and structure of the groundwater from the wells were analyzed via clonal libraries of partial SSU rRNA gene sequences, a phylogenetic chip array (Bacteria and Archaea), and a functional gene chip array over time. LIBSHUFF analysis for the clonal libraries of the re-circulating wells showed that over each phase of manipulation for uranium immobilization, the bacterial communities of the inner zones of biostimulation were more similar to each other than those of the outer protective zones. The outer protective zones were more similar to the injection well. LIBSHUFF analyses for the clonal libraries from FW-104 (injection), FW-101 and FW-102 (biostimulation) showed that bacterial communities of the three wells were initially similar but developed changes through time. FW-101

* Presenting author

and FW-102 bacterial communities developed changes in parallel, while those of FW-104 showed gradual changes. Diversity indices showed that bacterial diversity tended to increase during the initial phase of uranium bioreduction and decreased toward the end of uranium bioreduction (i.e., low U(VI) levels). In addition, when electron donor was added to the subsurface, community diversity increased with a subsequent decline in U(VI) levels. However, when levels of potential electron acceptors decreased, community diversity also decreased. As uranium levels declined, increasing *Desulfovibrio*- and *Geobacter*-like sequences were detected from the clonal libraries; moreover, *Desulfovibrio*-like sequences predominated over time. The results were further confirmed via RT-PCR, and RT-PCR results correlated with OTU and PhyloChip distributions for *Desulfovibrio*. PhyloChip analyses also demonstrated the presence and dynamics of both acetoclastic and hydrogenotrophic methanogens. The microbial community dynamics from one of the 4 frequently sampled monitoring wells (FW 102-3) was intensively analyzed with a functional gene array containing 27,000 probes covering 10,000 genes and >100 gene categories. The microarray data indicated that during the uranium reduction period, both FeRB and SRB populations reached their highest levels at Day 212, followed by a gradual decrease over 500 days. The uranium concentrations in the groundwater were significantly correlated with total abundance of c-type cytochrome genes ($r=0.73$, $p<0.05$) from *Geobacter*-type FeRB and *Desulfovibrio*-type SRB, and with the total abundance of *dsr*AB (dissimilatory sulfite reductase) genes ($r=0.88$, $p<0.05$). Mantel test of microarray data and chemical data also indicated that there was significant correlation between the differences of uranium concentrations and those of total c-cytochrome gene abundance ($r=0.75$, $p<0.001$) or *dsr*AB gene abundance ($r=0.72$, $p<0.01$). The changes of more than a dozen individual c-type cytochrome genes from *Geobacter sulfurreducens* and *Desulfovibrio desulfuricans* showed significant correlations to the changes of uranium concentrations among different time points. Also the changes of more than 10 *dsr*AB-containing populations, including both cultured (e.g. *Desulfovibrio* spp., *Desulfotomaculum*, and *Thermosedulfovibrio*) and non-cultured SRB were significantly related to the changes in uranium concentrations. These results suggested the importance of these functions for *in situ* uranium reduction. Interestingly, the changes of several *dsr*AB sequences previously recovered from this site (e.g., FW003269B, FW300181B) showed significant correlations to the changes in uranium levels. In conclusion, the microbial community composition and structure changed upon stimulating for uranium bioreduction conditions, and that sequences representative of the sulfate-reducers *Desulfovibrio* spp. and metal-reducers *Geobacter* spp. were detected in wells that displayed a decline in U(VI).

16 GTL

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

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

[1]Lawrence Berkeley National Laboratory, Berkeley, California; [2]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [3]Diversa, Inc., San Diego, California; [4]University of Missouri, Columbia, Missouri; [5]University of Washington, Seattle, Washington; [6]Miami University, Oxford, Ohio; [7]Genomatica, San Diego, California; [8]University of Oklahoma, Norman, Oklahoma; and [9]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics: GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

Field Studies

Environmental Characterizations. Clonal libraries for the SSU rRNA gene are a commonly used tool for the characterization of bacterial communities, and confidence intervals were predicted for accuracy of sequence determination from SSU rRNA libraries. The data and the model results suggested that similarity values below 0.995 are likely derived from dissimilar sequences at a confidence level of 0.95, and not sequencing errors. The results confirmed that screening by direct sequence determination could be reliably used to differentiate at the species level (Fields et al., 2006). Clonal libraries were then used to characterize changes in community structure along a contaminant plume (Oak Ridge, TN) in terms of phylogenetic, functional, and geochemical changes. Such studies are essential to understand how a microbial ecosystem responds to perturbations. Our results indicated that different gene sequences estimated different relationships between populations within the microbial communities. However, functional groups that respond differently under a particular perturbation should have different patterns of diversity along the contaminant gradient in relation to growth and competitive displacement, and the data supported this hypothesis (Fields et al., 2006b). An additional study characterized the community changes in a fluidized bed reactor for the treatment of uranium-contaminated groundwater. Changes in community structure and composition were correlated to operating conditions, and relationships between diversity and stability were discussed (Hwang et al., 2006). Our current work has been the identification of predominant populations in the uranium/nitrate-contaminated subsurface during bio-stimulation for heavy metal reduction. The data, thus far, indicated that when electron donor was added to the subsurface, community diversity increased with a subsequent decline in U(VI) levels. However, when levels of potential electron acceptors decreased, community diversity also decreased. As uranium levels declined, increasing *Desulfovibrio* and *Geo-*

* Presenting author

bacter-like sequences were detected from the clonal libraries; moreover, Desulfovibrio-like sequences predominated over time (Hwang et al., 2006b). 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. 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 stasis-induced genes and gene networks by looking at transition of *D. vulgaris* from exponential- to stationary-phase during electron donor depletion. Our results demonstrated that *D. vulgaris* cells altered gene expression profiles in response to carbon and energy depletion, and that gene expression during stationary-phase was not static. In particular, genes related to phage, carbohydrate flux, outer envelop, and iron homeostasis played a major role in the cellular response to nutrient deprivation under the tested growth conditions.

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

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

Genome Sequence. The genome sequence for *Desulfovibrio vulgaris* DePue strain DP4 has been closed and is now being annotated in collaboration with other ESPP investigators (see other abstract).

Dual culture systems. We achieved steady state growth of a syntrophic association between *Desulfovibrio vulgaris* and *Methanococcus maripaludis* in chemostats equipped with on-line monitoring of volatile metabolites (hydrogen, methane, CO_2). In association with other ESPP investigators, transcriptional analyses of replicated chemostat-grown cocultures and monocultures were completed (see other abstracts). Characterization of the corresponding proteomes is now in progress. In collaboration with the Wall laboratory, mutants in genes implicated in syntrophic growth were examined

for growth in monoculture and in coculture (see other abstracts). These experiments strongly suggest that the Coo Hydrogenase plays a central role in energy conservation during syntrophic growth, possibly functioning as a proton translocating hydrogenase. A second candidate energy converting hydrogenase, Ech, was demonstrated to play a minor role in syntrophic growth, but has been implicated in the production of reduced ferredoxin required for synthesis of pyruvate when growing on hydrogen and acetate.

Stress Experiments

High Throughput Biomass Production. Producing large quantities of high quality and defensibly reproducible cells that have been exposed to specific environmental stressors is critical to high throughput and concomitant analyses using transcriptomics, proteomics, metabolomics, and lipidomics. Culture of *D. vulgaris* is made even more difficult because it is an obligate anaerobe and sulfate reducer. For the past four years, our Genomics:GTL VIMSS project has developed defined media, stock culture handling, scale-up protocols, bioreactors, and cell harvesting protocols to maximize throughput for simultaneous sampling for lipidomics, transcriptomics, proteomics, and metabolomics. All cells for every experiment, for every analysis are within two subcultures of the original ATCC culture of *D. vulgaris*. In the past four years we have produced biomass for 120 (40 in the last year) integrated experiments (oxygen, NaCl, NO_3, NO_2, heat shock, cold shock, pH, Cr, and mutants Fur, Zur, Per, and MP(-)) each with as much as 200 liters of mid-log phase cells (3×10^8 cells/ml). This year new continuous culture extremophile bioreactors were brought online so that six reactors (1-3L) can be operated continuously. This enables U.S. to produce as much as 300L of mid-log phase anaerobe cells in 5 days. In addition, more than 80 adhoc experiments for supportive studies have been done each with 1-6 liters of culture. All cultures, all media components, all protocols, all analyses, all instruments, and all shipping records are completely documented using QA/QC level 1 for every experiment and made available to all investigators on the VIMSS Biofiles database (http://vimss.lbl.gov/perl/biofiles). To determine the optimal growth conditions and determine the minimum inhibitory concentration (MIC) of different stressors we adapted plate reader technology using Biolog and Omnilog readers using anaerobic bags and sealed plates. Since each well of the 96-well plate produces an automated growth curve, over more than 200 h, this has enabled U.S. to do more than 10,000 growth curves over the last three years. Since the Omnilog can monitor 50 plates at a time, this allows U.S. to do more than 5,000 growth curves in a year.

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

Synchrotron FTIR Spectromicroscopy for Real-Time Stress Analysis. This year we further the synchrotron FTIR Spectromicroscopy approach for studying roles of cellular compositions and physiological states during stress and adaptive responses in individual *D. vulgaris* triggered by air-level oxygen. Previously, the FTIR spectroscopy approach has allowed U.S. to detect *in situ* changes in intracellular molecules or molecular structures, and to nondestructively monitor and quantify metabolites produced in response to different stresses. This is because the chemical and structural information of molecules associated with cellular processes inside *D. vulgaris* are contained in each infrared spectrum; thus, one can extract chemical and structural information from each spectrum regarding the physiological conditions of a cell or a group of cells. The improved FTIR spectroscopy approach includes an additional molecular screening procedure, which allows U.S. to rapidly identify individual

* Presenting author

D. vulgaris cells that satisfy a targeted chemical composition and physiological state. Such rigidly controlled experimental conditions at chemical and biological levels would improve the reproducibility of experimental results. To date, we have evaluated the new FTIR approach in four different experimental systems using monolayers of wild-type *D. vulgaris* at the early stationary phase. For the first two systems, individual *D. vulgaris* cells of different compositions were maintained anaerobically, which have allowed U.S. to establish baselines for the molecular changes and the timescales associated with cellular processes during anaerobic metabolism. For the remaining two systems, individual *D. vulgaris* cells of different compositions were exposed to air-level oxygen, which have allowed U.S. to establish baselines for the molecular changes and the timescales associated with cellular processes during oxidative stress induced adaptive responses. Many of these results have been confirmed by analysis of microscopy images and biochemical essays. These studies will enable U.S. to do in depth studies of stress mechanisms with the new created mutants from the Functional Genomics Core of the project.

17 GTL

Microarrays + NanoSIMS: Linking Microbial Identity and Function

Jennifer Pett-Ridge[1]* (pettridge2@llnl.gov), Peter K. Weber,[1] Paul Hoeprich,[1] Philip Banda,[1] **Ian Hutcheon**,[1] Eoin Brodie,[2] and Gary Andersen[2]

[1]Lawrence Livermore National Laboratory, Livermore, California and [2]Lawrence Berkeley National Laboratory, Berkeley, California

Project Goals: We are using a high resolution ion microprobe (Nano Secondary Ion Mass Spectrometer NanoSIMS) to link microbial metabolism to molecular structures and produce a detailed view of how isotopically marked species propagate throughout individual cells. We expose microbes to stable isotope tracers and then map the tracer distribution with the NanoSIMS. Images of cells and microarrays reveal locations of active growth, nutrient fluxes between cells, and functional roles of community members.

In order to predict how microbes may react under given environmental conditions, or be engineered to perform useful functions, it is essential to understand the relationships between their molecular and metabolic profiles. Indeed, our need to understand both the identity and functional capacity of microorganisms is increasing as researchers seek to: a) understand spatial and metabolic relationships within complex microbial communities, b) exploit microbial traits for bioengineered fuel cells and cellulose conversion to biofuels, and c) utilize microbes to remediate contaminated sites.

We are addressing these goals by developing a new methodology, "NanoSIP", combining the power of re-designed oligonucleotide microarrays with nano-scale secondary ion mass spectrometry (NanoSIMS) analyses in order to link the identity of microbes to their functional roles. Building upon the concept of stable isotope probing (SIP) (Radajewski *et al.* 2000), we are isotopically labeling microbial nucleic acids by growing organisms on ^{13}C enriched substrates. When hybridized to a high density oligonucleotide microarray we can use the high spatial resolution and high sensitivity of the NanoSIMS to detect isotopic enrichment in ribosomal RNA fragments identified through fluorescent hybridization to a newly engineered oligonucleotide microarray. This approach will allow U.S. to directly link microbial identity and function.

The NanoSIMS is an imaging secondary ion mass spectrometer with the unprecedented combination of high spatial resolution, high sensitivity and high mass specificity. It has 50 nm lateral resolution and is capable of detecting 1 of every 200 carbon atoms in a sample while excluding isobaric interferences. We have previously used the NanoSIMS to document isotopic and elemental variations in tiny bioparticles such as *Bacilllus* spores, bacterial cells and lipid bilayers. Since the spot or feature size on a microarray is typically microns in diameter, and can contain millions of copies of an oligonucleotide probe, t the NanoSIMS has the detection capability to resolve array spots labeled with ^{12}C rRNA from those labeled with ^{13}C rRNA.

We are currently in the 'proof-of-concept' phase of method development and are testing the technique using pure cultures of ^{13}C-labelled microbes. Using environmental isolates from a tropical soil, we cultured 2 strains each of fungi, gm (+) bacteria, gm (-) bacteria and actinomycetes with ^{13}C-glucose. Cultures were repeatedly subsampled during exponential phase growth in order to generate a set of samples with a range of isotopic enrichments. We have extracted DNA from these isolates, sequenced the 16S/ITS region and generated 25-mer oligonucleotide probes for each organism. This probe set can be printed onto high density oligonucleotide microarrays using the NimbelGen synthesizer in the LLNL-Livermore Microarray Center (LMAC). The arrays we are using are newly engineered to have a more conductive surface and higher reproducibility relative to traditional glass/silane microarrays. These advances allow U.S. to successfully analyze microarray slides with a nano-secondary ion mass spectrometer (NanoSIMS), generating isotopic and elemental abundance images of the array surface, and indicating which organisms utilized the isotopically labeled substrate. We intend to apply the method to complex microbial communities found in biofilms and soils in the near future.

Reference

1. Radajewski S, Ineson P, Parekh NR & Murrell JC 2000. Stable-isotope probing as a tool in microbial ecology. **Nature** 403: 646-649

18 GTL

NanoSIMS Analyses of Molybdenum Indicate Nitrogenase and N-Fixation Activity in Diazotrophic Cyanobacteria

Jennifer Pett-Ridge,[1] Juliette Finzi,[2] **Ian D. Hutcheon**[1] (hutcheon1@llnl.gov), Doug Capone,[2] and Peter K. Weber[1]*

[1]Chemistry, Materials, and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California and [2]Department of Marine Biology, University of Southern California, Los Angeles, California

Project Goals: We are using a high resolution ion microprobe (Nano Secondary Ion Mass Spectrometer NanoSIMS) to link microbial metabolism to molecular structures and produce a detailed view of how isotopically marked species propagate throughout individual cells. We expose microbes to stable isotope tracers and then map the tracer distribution with the NanoSIMS. Images of cells and microarrays reveal locations of active growth, nutrient fluxes between cells, and functional roles of community members.

Diazotrophic cyanobacteria are capable of both CO_2 and N_2 fixation, yet must separate these two functions because the nitrogenase enzyme critical to N_2 fixation is inhibited by O_2 produced during

* Presenting author

photosynthesis. Some lineages, such as *Anabaena oscillarioides*, use specialized cells (heterocysts) to maintain functionwal segregation. However the mechanism of this segregation is poorly understood in the undifferentiated filamentous *Trichodesmium spp.*, an important component of marine primary production in the tropical and subtropical North Atlantic. While some research on *Trichodesmium IMS101* suggest a temporal segregation of the nitrogen and carbon fixing processes, others indicate nitrogen fixation is spatially isolated in differentiated cells called diazocytes (Fredriksson and Bergman 1997).

In order to isolate the intracellular location of N fixation in both species, we used a combination of TEM, SEM and NanoSIMS analysis to map the distribution of C, N and Mo (a critical nitrogenase co-factor) isotopes in intact cells. NanoSIMS is a powerful *in situ* analysis tool which combines nanometer-scale imaging resolution with the high sensitivity of mass spectrometry. Using cells grown in a $^{13}CO_2$ and $^{15}N_2$ enriched atmosphere, our analyses show that heterocysts in *Anabaena* have Mo concentrations four times higher than those of non-N-fixing vegetative cells. Recently fixed N does not accumulate at the site of fixation, but instead is quickly translocated to vegetative cells, presumably to fuel the demands of photosynthesis, storage and cell division.

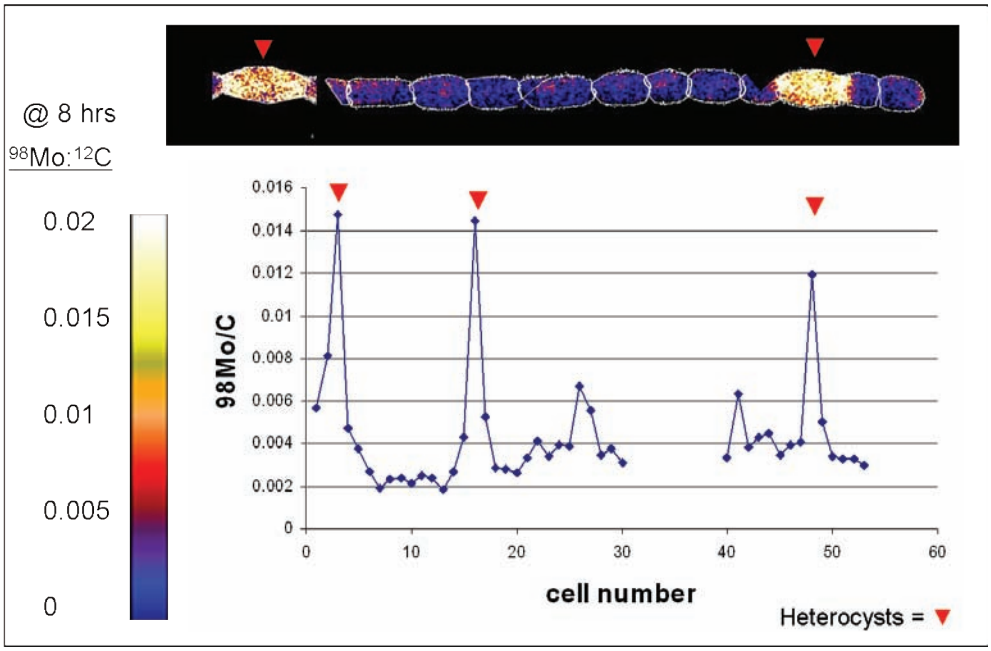


Figure. NanoSIMS image and dataplot of Mo concentrations in a strand of *Anabaena oscillarioides* heterocyst and vegetative cells grown under N-fixation conditions. Brighter colors indicate higher concentrations.

In the non-heterocystous *Trichodesmium IMS101*, Mo is concentrated in sub-regions of individual cells, and is not associated with regions of N storage (cyanophycin granules) which are clearly evident in lateral section TEM images. Average cellular concentrations of Mo increased from 1 (± 0.2) ppm to 86 (± 20) ppm during the same early afternoon period when a jump in ^{15}N (and thus N-fixation) was observed. Rare and randomly located cells contained extremely high concentrations of Mo (>2000 ppm).

We suggest that NanoSIMS mapping of metal enzyme co-factors is a powerful method of identifying physiological and morphological characteristics within individual bacterial cells, and could be

used to complement more traditional analyses such as immunogold labeling. Such combinations of NanoSIMS analysis and high resolution microscopy allow isotopic analysis to be linked to morphological features and hold great promise for fine-scale studies of bacteria metabolism.

Reference

1. Fredriksson, C. & Bergman, B. (1997). Ultrastructural characterization of cells specialized for nitrogen fixation in a non-heterocystous cyanobacterium, *Trichodesmium*. *Protoplasma* 197, 76–85

19 GTL

Application of a Novel Genomics Technology Platform

Mircea Podar,[1] Carl Abulencia,[2] Don Hutchinson,[2] Joseph Garcia,[2] Lauren Hauser,[1] Cheryl Kuske,[3] and **Martin Keller**[1]* (kellerm@ornl.gov)

[1]Bioscience Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [2]Diversa Corporation, San Diego, California; and [3]Los Alamos National Laboratory, Los Alamos, New Mexico

Project Goals: The Application of a Novel Genomics Technology Platform combines an isolation method based on fluorescence in situ hybridization (FISH) and cell sorting by flow cytometry, with whole genome amplification (MDA) to obtain a sufficient amount of DNA for sequencing whole genomes of uncultured microorganisms. Soil bacterial representatives of candidate division TM7 were specifically FISH-stained, in suspension, and isolated by flow cytometry. The genomic DNA was subsequently amplified by MDA for the construction of libraries for shotgun sequencing.

Application of cultivation-independent molecular phylogenetic approaches to study microbial communities in the environment led to the discovery of an unexpected genetic diversity and has been followed by an increasing repertoire of environmental genomic tools (expression microarrays, proteomics, and metabolomics). However, the cost and time effort for genomic characterization of most microbial communities through shotgun sequencing is prohibitive due to high microbial diversity and complex distributions of abundance and genome size for the constituent species. To overcome some of these hurdles, we propose a targeted genomic approach. This process combines an isolation method based on fluorescence in situ hybridization (FISH) and cell sorting by flow cytometry, with whole genome amplification (MDA) to obtain a sufficient amount of DNA for sequencing whole genomes of uncultured microorganisms. Soil bacterial representatives of candidate division TM7 were specifically FISH-stained, in suspension, and isolated by flow cytometry (manuscript in preparation). The genomic DNA was subsequently amplified by MDA for the construction of libraries for shotgun sequencing. Based on SSU rRNA sequences, the soil sample studied contained mostly members of the Proteobacteria (35%), Acidobacteria (38%), Gemmatimonadetes (16%) and, at much lower abundance levels (2% or less), representatives of several other phyla. Candidate division TM7 bacteria were among those low-abundance groups, which was appropriate for our goal of targeting a minor constituent of the community for genomic characterization.

We have targeted this approach to the TM7 using specific FISH-staining, in suspension, and isolation of stained bacteria by flow cytometry. A cellular fraction prepared from the soil sample was used for hybridization with a fluorescently labeled oligonucleotide specific for the TM7 phylum (TM7905). Using flow cytometry we detected a small fraction of cells (0.02%) which had a fluorescence level approximately 10 times higher than background based on the unstained control popula-

* Presenting author

tion. Fluorescence cells were sorted in pools of various sizes and used for chromosomal amplification. The selected genomic DNA was subsequently amplified by MDA for the creation of libraries for shotgun sequencing of whole genomes. Based on test experiments we determined that five was the fewest number of cells that balanced efficient genomic amplification with low levels of amplification artifacts and chimeric clones.

The MDA-amplified genomic DNA from five sorted cells was used as template for SSU rRNA gene amplification. Among the 69 sequences, 61 (89%) represented a TM7 bacterium. The remaining eight sequences were found to be nearly identical (>99.5%) to SSU ribosomal genes from several environmental *Pseudomonas* isolates including *P. rhodesiae*, an organism isolated from natural mineral waters. These clones may therefore represent an actual *Pseudomonas* cell that was sorted by flow cytometry from the soil sample rather than from contamination of the reagents or instruments.

Sequencing of the amplified DNA has resulted in identification of genes that are from the TM7 genome and will give insights to the functioning of this group. End sequences from12,000 clones were generated using Sanger-sequencing. After filtering out the low quality and obvious chimeric reads based on Phred/Phrap, approximately 20,000 reads were assembled into contigs using Phrap. Contigs that contained genes with high similarity values to known *Pseudomonas* genes had also elevated GC content (>54%) relative to the bulk of the sequences (<50%) and were filtered out as representing the contaminant. The remaining sequence data, representing ~600kb of, constitutes approximately 15-20% of the TM7 genome, based on statistical distribution of universally present bacterial genes. This genomic data allows for the first time detailed evolutionary analyses of the TM7 phylum as well as insight into the soil TM7 bacterial ecology and metabolism.

Research sponsored by the Genomics:GTL program, Office of Biological and Environmental Research, U.S. Department of Energy Grant No. DE-FG02-04ER63771

20 GTL

Genome-Scale Analysis of the Physiological State of *Geobacter* Species During *In Situ* Uranium Bioremediation

Dawn E. Holmes* (dholmes@microbio.umass.edu), Regina A. O'Neil, Milind A. Chavan, Muktak Aklujkar, and **Derek R. Lovley**

Department of Microbiology, University of Massachusetts, Amherst, Massachusetts

Project Goals: The overall goal of the Genomics:GTL *Geobacter* Project is to develop genome-based in silico models that can predict the growth and metabolism of *Geobacteraceae* under a variety of environmental conditions. These models are required in order to optimize practical applications of *Geobacteraceae* that are relevant to DOE interests. The goals of Subproject I and II are to determine the genetic potential of the *Geobacteraceae* present in subsurface environments, and to describe genome-wide patterns of gene expression in *Geobacteraceae* species in subsurface environments. This not only provides information on what metabolic modules need to be included in the in silico models but makes it possible to monitor the metabolic state and rates of metabolism in diverse environments by measuring transcript levels of key diagnostic genes.

The design of optimal bioremediation strategies for contaminated Department of Energy subsurface sites has been hindered by a lack of information on the *in situ* physiological state of the microorganisms involved in important bioremediation processes and the inability to predict how the microbial

community will respond to different amendments that might be made to promote bioremediation. It is now clear from numerous studies in multiple laboratories at a diversity of sites that stimulation of dissimilatory metal reduction to promote *in situ* bioremediation of uranium and other contaminant metals frequently results in the emergence of *Geobacter* species as the dominant metal-reducing microorganisms. *Geobacter* species are also the predominant organisms involved in the oxidation of various organic contaminants coupled to the reduction of the Fe(III) oxides that are naturally abundant in most subsurface environments.

Further analysis of the *Geobacter* species that predominate in a diversity of subsurface environments revealed a 'subsurface clade' of *Geobacter* species that are consistently the dominant *Geobacter* in a geographic and geochemical diversity of subsurface environments regardless of whether metal or organic contaminants are undergoing bioremediation. This finding greatly simplifies both the identification of gene target sequences for evaluation of the *in situ* metabolic state of *Geobacter* species during groundwater bioremediation and the development of genome-based *in silico* models to predict the metabolic and growth responses of *Geobacter* species under different potential bioremediation strategies.

Three independent and complementary approaches were taken in order to learn more about the genetic potential of the 'subsurface clade' of *Geobacter* species: 1) small insert libraries of genomic DNA extracted directly from the environment were sequenced; 2) genomic DNA was amplified from single cells recovered from the subsurface and sequenced; and 3) strains of *Geobacter* with 16S rRNA gene sequences identical or highly similar to the sequences that predominate during bioremediation were recovered in pure culture, and their genomes were sequenced. Approaches 1 and 2 provide information on the genotypic potential of the *Geobacter* species that predominate during subsurface bioremediation, but are limited in value because: 1) many of the genes identified are of unknown function or if they have an annotated function, their physiological role in *Geobacter* is unclear; 2) it is not possible to predict patterns of gene expression from sequence data alone; and 3) many of the most basic and important phenotypic characteristics, such as optimal conditions for growth, growth yields, etc. cannot yet be predicted from genome sequences alone. Thus, the ability to conduct genome scale analysis of the physiology of environmentally relevant isolates is key to understanding *in situ* physiology and the development of predictive *in silico* models.

Genomes of multiple subsurface *Geobacter* isolates have been sequenced or will be completed shortly. These include: *G. uraniumreducens*, *Geobacter* species M21, and *Geobacter* species M18 recovered from *in situ* uranium bioremediation experiments at the DOE-ERSP field study site in Rifle, CO; strain FRC-32, a *Geobacter* species recovered from an *in* situ uranium bioremediation experiment at the DOE-ERSP Field Research Center at Oak Ridge National Laboratories; *Geobacter* strains Ply1 and Ply4 which were recovered from an acetate-impacted aquifer that serves as an analog for long-term *in situ* uranium bioremediation; and *G. bemidjiensis*, recovered from the Fe(III)-reducing zone of a petroleum-contaminated aquifer. Preliminary results show substantial similarities in the genome sequences of these isolates and the genome sequences obtained from sequencing genomic DNA extracted from the subsurface.

In order to learn more about the physiology of *Geobacter* species growing in subsurface sediments pure cultures of *Geobacter* species were inoculated into sterilized, uranium-contaminated sediments from the ERSP study site in Rifle, CO and the sediments were amended with acetate to simulate conditions during *in situ* uranium bioremediation. Compared to growth on soluble electron acceptors, all three of the species examined, *G. uraniumreducens*, *G. sulfurreducens*, and *G. metallireducens*, had significant increases in transcripts for multiple genes for *c*-type cytochromes, not only during growth in sediment but also in culture medium when Fe(III) or Mn(IV) oxides served as the electron acceptor. There was also increased expression of genes for multicopper oxidase proteins.

* Presenting author

The within-strain similarity in gene expression with all three electron acceptors suggests that the mechanisms for electron transfer to subsurface sediment oxides and oxides prepared in the laboratory to simulate sediment oxides are similar. However, between species there were substantial differences in which cytochrome genes were most highly expressed, reflecting the lack of cytochrome gene conservation in *Geobacter* species.

In contrast to the lack of conservation of cytochrome genes, there is high conservation of many other genes across *Geobacter* species and for these genes there were highly similar expression patterns. For example, a number of genes that encode proteins involved in chemotaxis and motility and phosphorus limitation were significantly up-regulated in all of the organisms during growth in sediments or on Fe(III) or Mn(IV) oxides. Furthermore, genes encoding proteins involved in nitrogen fixation, heavy metal stress, and oxidative stress were up-regulated in all three species during growth in sediments, but not when Fe(III) oxide or Mn(IV) oxide were provided as the electron acceptor.

Remarkably, gene expression patterns of pure cultures grown in sediments were highly similar to the *in situ* gene expression of the *Geobacter* species that predominated during *in situ* uranium bioremediation at the Rifle study site. The *Geobacter* species in the groundwater had high transcript levels for genes involved not only in electron transfer to Fe(III) oxides, but also chemotaxis, motility, phosphorus uptake, nitrogen fixation, heavy metal stress, and oxidative stress. These results demonstrate that it is possible to reliably monitor the metabolic state of *Geobacter* species involved in *in situ* uranium bioremediation and suggest that detailed, genome-based physiological studies with pure cultures of environmentally relevant *Geobacter*species can provide insight into the physiology of *Geobacter* species living in subsurface environments. This has important implications for the ability of *in silico* models developed from pure cultures to predict growth and metabolism under different conditions in the subsurface.

Section 3
Protein Production and Characterization

21 GTL
High Throughput Selection of Affinity Reagents

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

Los Alamos National Laboratory, Los Alamos, New Mexico

Project Goals: The goals of this GTL funded project are to implement a high throughput selection and screening system for affinity reagents, with the ability to select against proteins and post-translational modifications. This has required re-engineering of the whole selection and screening process, described in the displayed posters.

Antibodies are the most widely used binding ligands in research. However, they suffer from a number of problems, especially when used in molecular diversity techniques. These include low expression levels, instability and poor cytoplasmic expression, as well the inability to detect binding without the use of secondary reagents. In this project we are creating an integrated in vitro system which will

allow U.S. to select affinity reagents against proteins of interest on a genomic scale. This has involved re-engineering of the whole selection and screening process. Within this context we have developed 1) novel affinity reagents based on fluorescent proteins which resolve many of the problems associated with antibodies (1, 2); 2) new selection methods for such fluorescent affinity reagents (3, 4); 3) high throughput screening systems using flow cytometry (5); 4) eliminated the need for helper phage in phage display selections (6), and 5) shown the application of some of these methods to the selection of antibodies recognizing post-translation modifications independently of sequence context (7)

References

1. Dai, M., Fisher, H.E., Temirov, J., Kiss, C., Phipps, M.E., Pavlik, P., Werner, J.H. and **Bradbury, A.R.M.** (2006) The creation of a novel fluorescent protein by guided consensus engineering, Prot. Eng. Design Selection In press
2. Kiss, C., Fisher, H. Pesavento, E., Dai, M., Valero, R., Ovecka, M., Nolan, R., Phipps, L., Velappan, N., Chasteen, L., Martinez, J., Waldo, G.S., Pavlik, P. and **Bradbury, A.R.M.** (2006) Antibody binding loop insertions as diversity elements. Nuc. Acids Res., **34**, e132
3. Dai, M., Pavlik, P. and **Bradbury, A.R.M.** (2007) Using T7 phage display to select GFP based binders, in preparation
4. Velappan, N., Fisher, H., Kiss, C., Chasteen, L., Pavlik, P. and Bradbury, A.R.R. (2007) Optimizing export signals for the phage display of cytoplasmic proteins, in preparation
5. Ayriss, J., Woods, T., **Bradbury, A.R.M.** and Pavlik, P. (2006) High throughput screening of single chain antibodies using multiplexed flow cytometry, J. Proteomic Res. In press
6. Chasteen, L., Ayriss, J., Pavlik, P. and **Bradbury, A.R.M.** (2006) Eliminating helper phage from phage display, Nuc. Acids Res., **34**, e145
7. Kehoe, J.W., Velappan, N., Walbolt, M., Rasmussen, J., King, D., Lou, J., Knopp, K., Pavlik, P., Marks, J.D., Bertozzi, C.R., and **Bradbury, A.R.M.** (2006) Using phage display to select antibodies recognizing post-translational modifications independently of sequence context. Molecular Cellular Proteomics, in press.

22 GTL

Progress on Fluorobodies

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

Los Alamos National Laboratory, Los Alamos, New Mexico

Project Goals: The goals of this GTL funded project are to implement a high throughput selection and screening system for affinity reagents, with the ability to select against proteins and post-translational modifications. This has required re-engineering of the whole selection and screening process, described in the displayed posters.

Antibodies are the most widely used binding ligands in research. However, they suffer from a number of problems, especially when used in molecular diversity techniques. These include low expression levels, instability and poor cytoplasmic expression, as well the inability to detect binding without the use of secondary reagents. We have developed novel affinity reagents based on fluorescent proteins which resolve many of these problems. However, selection of such affinity reagents remains problematic, because they do not appear to be very well displayed on phage. In addition to using a cytoplasmic phage, T7 (1), we have undertaken an examination of the ability of three different translocation

* Presenting author

pathways (Sec, SRP, TAT) used by *E. coli* to transfer proteins into the periplasm, to incorporate GFP and modified GFP into phage particles, by placing different leaders upstream of GFP (2). We find that while superfolder GFP is efficiently translocated and incorporated into phage in a functional manner with all three leaders, GFP which has been modified, by the insertion of a binding loop, for example, can only be effectively incorporated into phage using TAT based leaders. This provides an effective phage display platform with which to select fluorescent protein based affinity reagents.

Additional data on the success in selecting affinity reagents with intrinsic fluorescence will be presented.

References

1. Dai, M., Pavlik, P. and **Bradbury, A.R.M.** (2007) Using T7 phage display to select GFP based binders, in preparation
2. Velappan, N., Fisher, H., Kiss, C., Chasteen, L., Pavlik, P. and Bradbury, A.R.R. (2007) Optimizing export signals for the phage display of cytoplasmic proteins, in preparation

23 GTL

High Throughput Screening of Affinity Reagents: Eliminating Helper Phage from Phage Display by the Use of Helper Plasmids

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

Los Alamos National Laboratory, Los Alamos, New Mexico

Project Goals: The goal of this GTL funded project is the implementation of a high throughput affinity reagent selection and screening program against proteins and post-translational modifications. This has required re-engineering many aspects of the selection and screening process described in the displayed posters.

Phage display is a relatively straightforward technology used to generate binding ligands against a vast number of different targets, involvings the display of proteins or peptides, as coat protein fusions, on the surface of a phage or phagemid particles. However, the need to use helper phage for the replication and assembly of phagemid particles, during library production and biopanning, has prevented full automation of the selection process. Helper phage are added at precise periods of bacterial growth, and it is impossible to avoid contamination of the phage output with helper phage. We have eliminated the need to add helper phage by using "bacterial packaging cell lines" that provide the same functions. These cell lines contain M13 based helper plasmids that express phage packaging proteins which assemble phagemid particles as efficiently as helper phage, but without helper phage contamination; resulting in genetically pure phagemid particle preparations. Furthermore, by using constructs differing in the form of gene 3 that they contain, we have shown that the display, from a single library, can be modulated between monovalent (phagemid-like) to multivalent display (phage-like) without any further engineering. These packaging cells eliminate the use of helper phage from phagemid based selection protocols; reducing the amount of technical preparation, facilitating automation, optimizing selections by matching display levels to diversity, and effectively using the packaged phagemid particles as means to transfer genetic information at an efficiency approaching 100%.

By eliminating the need to add helper phage at precise stages of bacterial growth, and avoiding contamination of the output phagemid particles with helper phage, the use of these cells rather than helper phage will considerably facilitate automation of phage display selection.

Reference

1. Chasteen, L., Ayriss, J., Pavlik, P. and **Bradbury, A.R.M.** (2006) Eliminating helper phage from phage display, Nuc. Acids Res., **34**, e145

24 GTL

Selecting Affinity Reagents which Recognize Specific Post-Translational Modifications Independently of Sequence Context: The Sulfotyrosine Example

John Kehoe,[3] Jytte Rasmussen,[2] Monica Walbolt,[2] Jianlong Lou,[4] James D. Marks,[4] Peter Pavlik,[1] Carolyn Bertozzi,[2] and **Andrew Bradbury**[1]* (amb@lanl.gov)

[1]Los Alamos National Laboratory, Los Alamos, New Mexico; [2]University of California, Berkeley, California; [3]Centocor, Horsham, Pennsylvania; and [4]University of California, San Francisco, California

Project Goals: The goal of this GTL funded project is the implementation of a high throughput affinity reagent selection and screening program against proteins and post-translational modifications. This has required re-engineering many aspects of the selection and screening process described in the displayed posters.

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

We designed a number of different selection strategies, using peptides containing the sulfotyrosine modification as positive selectors in the presence of an excess of the non-modified peptide as blocking agent. We screened almost eight thousand clones after two or three rounds of selection and identified a single scFv able to recognize tyrosine sulfate in multiple sequence contexts. Further analysis shows that this scFv is also able to recognize naturally sulfated proteins in a sulfation dependent fashion, and its binding could be inhibited by soluble tyrosine sulfate, but not tyrosine or tyrosine phosphate, providing an excellent way to control for the specificity of binding. This scFv was converted into a full length IgG and into an scFv-AP fusion, both of which increased the stability. This antibody has been distributed to a number of different groups which have used it successfully, some results of which will be presented.

It has proved to be extremely difficult to generate antibodies able to recognize post-translational modifications independently of sequence context by immunization, with antibodies against phosphotyrosine being the only well documented example. The use of phage display, as described here,

* Presenting author

provides proof of principle for the use of this technology to develop similar reagents against other post-translational modifications.

Reference

1. Kehoe, J.W., Velappan, N., Walbolt, M., Rasmussen, J., King, D., Lou, J., Knopp, K., Pavlik, P., Marks, J.D., Bertozzi, C.R., and **Bradbury, A.R.M.** (2006) Using phage display to select antibodies recognizing post-translational modifications independently of sequence context. Molecular Cellular Proteomics, in press.

25 GTL

A Total Chemical Synthesis Approach to Protein Structure and Function

Stephen Kent* (skent@uchicago.edu), Duhee Bang, Thomas Durek, Zachary Gates, Erik Johnson, Brad Pentelute, and Vladimir Torbeev

Institute for Biophysical Dynamics, University of Chicago, Chicago, Illinois

Project Goals: Our goal is to address the known limitations of chemical protein synthesis, based on our intimate understanding of the current state of the art. Emphasis will be on the development of simple methods using low cost hardware wherever possible. In this way, we will develop a practical chemical protein synthesis technology applicable to the rapid preparation of milligram amounts of small and integral membrane protein targets based on predicted gene sequence data. We will prototype the application of these methods to selected proteins of the model organism *Shewanella oneidensis* and proteins from a range of other sources, to illustrate potential application of chemical protein synthesis to validating the annotation of microbial genomes. The resulting knowledge will form the basis for future high throughput, parallel chemical synthesis of protein molecules that are difficult to prepare by recombinant DNA expression methods.

Microbial 'proteins' are being discovered at an accelerating pace, thanks to the successes of genome sequencing. Using advanced bioinformatics, in the past ten years many tens-of-thousands of predicted proteins have been added to the databases. Our next challenge is to validate the annotation of microbial genomes in terms of the mature protein translation products and their putative functions. One powerful, if underappreciated, way of doing this is *total chemical protein synthesis* - the use of organic chemistry to construct the predicted polypeptide chain, followed by folding of the synthetic polypeptide to give the unique, defined tertiary structure of the protein molecule. The synthetic product is then used to confirm the predicted biochemical function. Proteins prepared by total chemical synthesis have proved to be especially useful for determining the three-dimensional structure of the protein molecule by high resolution X-ray crystallography. Subsequent to these baseline observations, total chemical synthesis provides an efficient and versatile tool for elucidating in unique ways the molecular basis of protein function. For example, variant synthetic proteins corresponding to predicted post-translational modifications (e.g. phosphorylation; lipidation) can be readily prepared in defined, metabolically stable forms and then used to explore the effects on biochemical function (Ref SEP). Chemical protein synthesis is uniquely enabling for the application to proteins of advanced biophysical methods: e.g. selective labeling with nmr probe nuclei; single molecule fluorescence studies.

Over the past ten years, many hundreds of protein molecules have been successfully prepared by total chemical synthesis, typically in multiple tens-of-milligram amounts of high purity, correctly folded product. Total synthesis is particularly suited to the efficient preparation of small proteins (less than ~100 residues), Cys-rich proteins, and integral membrane (IM) proteins. Modern total protein synthesis has evolved from the 'chemical ligation' methods introduced in the mid-1990s (Refs.). Unprotected synthetic peptide segments, spanning the amino acid sequence of the target polypeptide chain, are covalently joined to one another in quantitative yield, without enzymes, by chemoselective reaction of unique, mutually reactive functional groups on each segment. Native chemical ligation ('NCL'), thioester-mediated chemoselective reaction at Cys residues, is the most robust and useful of ligation chemistry developed to date. Chemical protein synthesis is straightforward and the outcome quite predictable; the challenge for most laboratories is making the peptide-thioester building blocks.

We will present a series of case studies from our ongoing work, to illustrate the current capabilities of chemical protein synthesis and some of its applications. These case studies include:

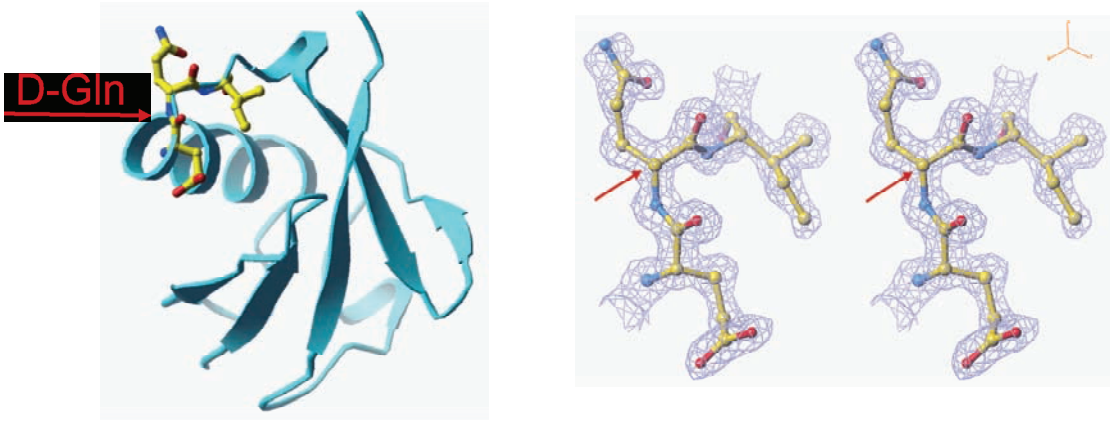


Figure 1. Synthesis and X-ray structures of ubiquitin and D-amino acid ubquitin analogues.

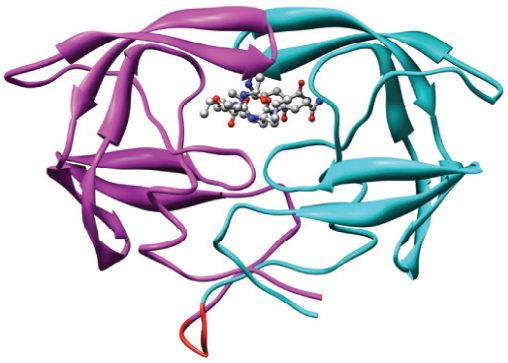


Figure 2. Convergent synthesis and Xray structure of a 203 amino acid residue 'covalent dimer' form of the HIV-1 protease

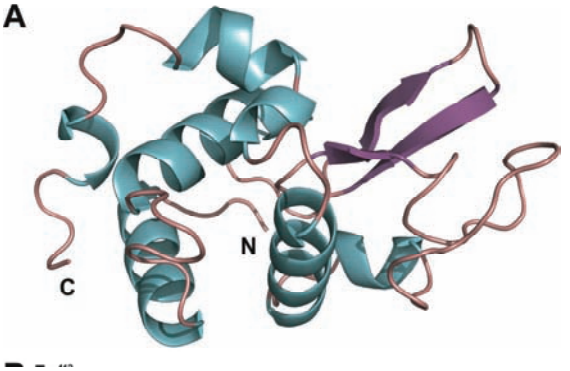


Figure 3. Convergent synthesis and 1.04Å Xray structure (above) of human lysozyme.

Other topics that will be discussed include: high efficiency synthesis of a series of transmembrane peptide-thioesters spanning the sequence of the protein diacylglycerol kinase, an integral membrane enzyme; 'kinetically controlled ligation' for the fully convergent synthesis of protein molecules; and, convergent synthesis of the plant protein crambin.

* Presenting author

We will describe recent innovations that extend the range of targets to which chemical protein synthesis can be applied. Future developments will include the high volume production of peptide-thioesters (modified chemistry, automated/parallel synthesis & purification), and high throughput micro-scale chemical protein synthesis using laboratory automation and microfluidics. Such improvements will enable the broad application of total chemical protein synthesis to the annotation of microbial genomes.

References

1. Constructing proteins by dovetailing unprotected synthetic peptides: backbone engineered HIV protease. M. Schnölzer, S. Kent *Science*, **256**, 221-225 (1992)
2. Synthesis of proteins by native chemical ligation. Philip E. Dawson, Tom W. Muir, Ian Clark-Lewis, Stephen B.H. Kent, *Science*, **266**, 776-779 (1994)
3. Synthesis of native proteins by chemical ligation. Dawson, P.E., Kent S.B.H. *Ann. Rev. Biochem*. **69**, 925-962 (2000)
4. Design and chemical synthesis of a homogeneous polymer-modified erythropoiesis protein. Gerd G. Kochendoerfer, et al., *Science*, **299**, 884-887 (2003)
5. Dissecting the energetics of protein α-helix C-cap termination through chemical protein synthesis. Duhee Bang, Alexey V. Gribenko, Valentina Tereshko, Anthony A. Kossiakoff, Stephen B. Kent*, George I. Makhatadze*, *Nature Chemical Biology*, **2**, 139-43 (2006)
6. Towards the total chemical synthesis of integral membrane proteins: a general method for the synthesis of hydrophobic peptide-αthioester building blocks. E.C.B. Johnson, S.B.H. Kent, *Tetrahedron Letters*, submitted (2006)
7. Kinetically-controlled ligation for the convergent chemical synthesis of proteins. Duhee Bang, Brad Pentelute, Stephen B.H. Kent, *Angew Chem Int Ed Engl.*, **45**, 3985-3988 (2006)
8. Total synthesis of proteins by convergent chemical ligation of unprotected peptides. T. Durek, D. J. Boerema, Z. P. Gates, S. Liu, B. L. Pentelute, V.Yu. Torbeev, Stephen B. H. Kent, submitted (2006)
9. Convergent chemical synthesis and high resolution X-ray structure of human lysozyme. Thomas Durek, Vladimir Yu. Torbeev, Stephen B. H. Kent, submitted (2006)
10. Convergent chemical synthesis and crystal structure of a 203 amino acid 'covalent dimer' HIV-1 protease enzyme molecule. Vladimir Yu. Torbeev, Stephen B. H. Kent, *Angew Chem Int Ed Engl*, accepted for publication (2006)

26 GTL

A Combined Informatics and Experimental Strategy for Improving Protein Expression

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

Center for Advanced Research in Biotechnology, Rockville, Maryland

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

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

Informatics methods have been used to examine a wide range of factors potentially affecting soluble expression, including protein family size, native expression level, low complexity sequence, open reading frame validity, amyloid propensity and inherent disorder. Of these, the most significant ones affecting expression outcome are native expression level, family size, and inherent disorder. Surprisingly, a relatively high fraction of disorder is also found to be a characteristic of 'singletons'. We are currently experimenting with machine learning methods, incorporating all of the above factors, as a means of predicting soluble expression.

Transcriptional profiling has revealed a reproducible pattern of gene expression in response to the accumulation of insoluble recombinant protein. The transcriptome partially overlaps those observed during heat shock induction or culture saturation, indicative of regulation, in part, by sigma factors 32 and 38 (encoded by rpoH and rpoS, respectively). We have used this information to develop a GFP reporter plasmid for insoluble protein accumulation, and identified sigma38 as a key regulator of its expression. Currently, efforts are underway to engineer the promoter of the reporter plasmid to decrease background GFP expression while retaining the ability to discriminate between soluble vs. insoluble protein accumulation.

Protein stability measurements on a set of 12 bacterial proteins have been performed using differential scanning calorimetry and chemical denaturation with guanidine hydrochloride. The data from the two methods are in good agreement, and confirm the earlier finding of that stability is not major factor in determining soluble expression. Work is now underway to investigate the folding properties of these proteins.

This project is supported by Genomics:GTL award DE-FG02-04ER63787.

27 GTL

Structural and Functional Characterization of a Periplasmic Sensor Domain from *Geobacter sulfurreducens* Chemotaxis Protein: A Novel Structure from a Family of Sensors in *Geobacteraceae*

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

[1]Biosciences Division, Argonne National Laboratory, Argonne, Illinois; [2]Requimte, CQFB, Dep. Quimica, FCT-UNL, Caparica, Portugal; and [3]Instituto de Tecnologia Quimica e Biologica, UNL, Oeiras, Portugal.

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

* Presenting author

includes modeling of structures based on their amino acid sequences, determination of their structures, and the functional interpretation of the structures, such as active sites and surface properties.

Geobacter sulfurreducens encodes over 100 cytochromes containing *c*-type hemes. *G. sulfurreducens* also has one of the largest numbers of proteins annotated as parts of the two-component signal transduction and/or chemotaxis pathways. Ten of the signal transducers have a periplasmic sensor domain which are homologous to each other, and contain sequence signature for *c*-type hemes (1). Two of these domains from methyl-accepting chemotaxis proteins encoded by genes GSU0582 and GSU0935, were expressed in *E. coli* co-transformed with the plasmid bearing cytochrome *c* maturation genes. The domains have about 135 residues, 40% of which are identical.

The heme groups in both proteins are five coordinated in their oxidized state and six-coordinated in their reduced state. The binding patterns for NO, and CO were determined by UV-Vis and NMR spectroscopies. Both proteins bind NO in their oxidized and reduced forms. CO only binds in the reduced state, replacing the endogenous sixth axial ligand of the heme. UV-Vis spectroscopy showed that imidazole is bound only in the oxidized state and it forms the sixth ligand to the heme. The ligand switch upon binding CO suggests a conformational change in the protein which could be a mechanism for signal transduction by these molecules. Both domains have a negative reduction potential: -169mV and -264mV for GSU0582 and GSU0935, respectively. The 95mV difference between their redox potentials suggests different biological functions for these domains.

Remarkably, although the UV-Vis spectra indicate that the heme of these domains is similar to that of cytochrome *c*', their structure is predicted by the program 3D-PSSM (2) to be homologous to CitAP, the periplasmic citrate-binding PAS domain of sensor kinase that does not contain heme (3). We now crystallized the sensor domain from of GSU0935 and determined its structure *de novo* using the anomalous dispersion of the iron atom of the heme at the Structural Biology Center beam line of the APS. As predicted by the program 3D-PSSM, the structure is indeed homologous to CitAP. Interestingly, only 13% of the residues is identical between CitAP and sensor domain of GSU0935; the heme binding site is found to be located in an inserted segment as predicted (1). The crystallographic refinement is in progress; details of the structure will be discussed.

The structure of sensor domain of GSU0935 is the first structure of a PAS domain that contains a covalently bound heme. This sensor domain from chemotaxis protein GSU0935 represents a previously unreported family of PAS-type periplasmic sensor domains; these domains could be part of an important mechanism for sensing redox potential or small ligands in the periplasm. Homologs to the sensor domains we identified in *G. sulfurreducens* are observed in various bacteria although they occur in larger numbers in the *Geobacteraceae.*

References

1. Londer YY, Dementieva IS, D'Ausilio CA, Pokkuluri PR & Schiffer M (2006) Characterization of a *c*-type heme containing PAS sensor domain from *Geobacter sulfurreducens* representing a novel family of periplasmic sensors in *Geobacteraceae* and other bacteria. *FEMS Microbiol Lett* **258**: 173-181.
2. Kelley LA, MacCallum RM & Sternberg MJE (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol* **299**: 499-520.
3. Reinelt S, Hofmann E, Gerharz T, Bott M & Madden DR (2003) The structure of the periplasmic ligand-binding domain of the sensor kinase CitA reveals the first extracellular PAS domain. *J Biol Chem* **278**: 39189-39196.

28 GTL

High-Throughput Production and Analyses of Purified Proteins

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

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

Project Goals: The work is aimed at improving the efficiency of high-throughput protein production from cloned coding sequences and developing a capacity for high-throughput biophysical characterization of the proteins obtained.

This work is aimed at improving the efficiency of high-throughput protein production from cloned coding sequences and developing a capacity for high-throughput biophysical characterization of the proteins obtained. Proteins are produced in the T7 expression system in *Escherichia coli*, which is capable of expressing a wide range of proteins. New vector/host combinations, combined with non-inducing and auto-inducing growth media, provide stable, reliable and convenient expression, even for proteins that are highly toxic to the host and cannot be maintained in the usual pET vectors.

Proteins produced from clones are often improperly folded or insoluble. Many such proteins can be solubilized and properly folded, whereas others appear soluble but remain aggregated or improperly folded. As high-throughput production of purified proteins becomes implemented in GTL projects and facilities, reliable analyses of the state of purified proteins will become increasingly important for quality assurance and to contribute functional information. Beam lines at the National Synchrotron Light Source analyze proteins by small-angle X-ray scattering (SAXS) to determine size and shape, X-ray fluorescence microprobe to identify bound metals, and Fourier transform infrared (FTIR) and UV circular dichroism (CD) spectroscopy to assess secondary structure and possible intermolecular orientation. A liquid-handling robot for automated loading of samples from 96-well plates for analysis at each of these stations has been built and implemented with purified proteins. These data are being used as a training set for neural network analysis of new proteins, to determine whether they are folded properly, obtain information on dynamics and stability, and provide an approximate structure classification.

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

* Presenting author

Section 4

Molecular Interactions

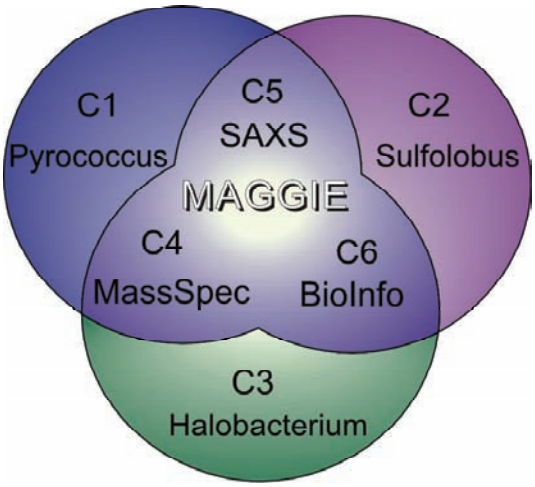
29 GTL

Molecular Assemblies, Genes, and Genomics Integrated Efficiently: MAGGIE

John A. Tainer* (jat@scripps.edu)

Life Science Division, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California

Project Goals: MAGGIE integrates an interdisciplinary team at Lawrence Berkeley National Lab with researchers at The Scripps Research Institute, the University of Georgia, the University of California Berkeley, and the Institute for Systems Biology into a unified Genomics GTL program. Major overall goals are 1) to facilitate instrument and technology development and optimizations though cross-disciplinary collaborations, 2) to comprehensively characterize complex molecular machines including protein complexes (PCs) and modified proteins (MPs) and 3) to provide critical enabling technologies and a prototypical map of PCs and MPs for the GTL Program.



MAGGIE integrates an interdisciplinary team at Lawrence Berkeley National Lab with researchers at The Scripps Research Institute, the University of Georgia, the University of California Berkeley, and the Institute for Systems Biology into a unified Genomics GTL program. Major overall goals are 1) to facilitate instrument and technology development and optimizations though cross-disciplinary collaborations, 2) to comprehensively characterize complex molecular machines including protein complexes (PCs) and modified proteins (MPs) and 3) to provide critical enabling technologies and a prototypical map of PCs and MPs for the GTL Program.

MAGGIE focuses on providing an integrated, multi-disciplinary program and synchrotron facilities at the Advanced Light Source (ALS) to achieve efficient key technologies and databases for the molecular-level understanding of the dynamic macromolecular machines that underlie all of microbial cell biology. Together the six MAGGIE Component Subprojects have complementary and synergistic capabilities that unite and leverage the biophysical strengths at LBNL and the ALS with those of top university and research institutes. The Program management and data sharing is promoting synergistic investigator interactions to provide interdisciplinary expertise and scientific critical mass to meet the emerging experimental challenges. Although a new program, we have already had substantial progress as shown on our website: http://masspec.scripps.edu/MAGGIE/index.php and in our publications (see below).

MAGGIE is moving to meet the challenges posed by comprehensive characterizations of molecular machines by combining the advantages of specific microbial systems with those of advanced technologies. We highlight 7 initial accomplishments for the overall program: 1) the *Pyrococcus* system is providing PCs and MPs from native biomass, 2) the *Sulfolobus* system is providing genetics for tagged complexes, 3) the *Halobacterium* system is providing extensive system biology results and capabilities, 4) novel developments in high throughput mass spectrometry promise to make large impacts on the research community, 5) the SIBLYS beamline and SAXS facilities are now working as unique and productive world class facilities to visualize PCs and MPs in solution, 6) graph theory is providing characterizations of protein module interactions using cliques, and 7) GAGGLE software is providing a superb technology for communications across multiple databases.

Publications from MAGGIE funding

Facciotti M.T., Pan M., Kaur A. Vuthoori M., Reiss D.J., Bonneau R., Shannon P., Srivastava A., Donahoe S.M., Hood L., Baliga N.S. "Structure of a general transcription factor specified global gene regulatory network," submitted, 2006

Schmid A.K., Reiss D.J., Kaur A., Pan M., King N., Hohmann L., Baliga N.S. "Tracking transcriptome and proteome dynamics during oxic/anoxic transitions in cellular physiology," submitted, 2006

Whitehead K., Kish A., Pan M., Kaur A., Reiss D.J., King N., Hohmann L., DiRuggiero J., Baliga N.S. "An integrated systems approach for understanding cellular responses to gamma radiation," *Mol Syst Biol*, 2: 47, 2006.

Schmid A., Baliga N. "Prokaryotic Systems Biology**,"** *In Cell Engineering*, El-Rubeai, M. (ed): Springer, 5, 2006.

Bonneau R., Reiss D.J., Shannon P., Facciotti M., Hood L., Baliga N.S., Thorsson V. "The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo," *Genome Biol*, 7: R36, 2006.

Reiss D.J., Baliga N.S., Bonneau R. "Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks," *BMC Bioinformatics*, **2006**, 7: 280.

Shannon P., Reiss D.J., Bonneau R., Baliga N.S. "Gaggle: An open-source software system for integrating bioinformatics software and data sources," *BMC Bioinformatics*, 7: 176, 2006.

Kaur A., Pan M., Meislin M., Facciotti M.T., El-Geweley R., Baliga N.S. "A systems view of haloarchaeal strategies to withstand stress from transition metals," *Genome Res*, 16: 841-854, 2006.

Want E.J., Nordstrom A., Morita H., Siuzdak G. "From Exogenous to Endogenous: The Inevitable Imprint of Mass Spectrometry in Metabolomics," submitted, 2006.

Go E.P., Wikoff W., Shen Z., O'Maille G., Morita H., Conrads T.P., Nordstrom A., Trauger S.A., Uritboonthai W., Lucas D., Chan K.C., Veenstra T.D., Lewicki H., Oldstone M.B., Schneemann A., Siuzdak G. "Mass Spectrometry Reveals Specific and Global Molecular Transformations during Viral Infection," *Journal of Proteome Research*, in press, 2006.

Shen Z., Want E.J., Chen W., Keating W, Nussbaumer W, Moore R, Gentle T.M., Siuzdak G. "Sepsis Plasma Protein Profiling with Immunodepletion, Three-Dimensional Liquid Chromatography Tandem Mass Spectrometry and Spectrum Counting," *Journal of Proteome Research*, in press, 2006.

* Presenting author

Northen T.R., Northen M.T., Nordstrom A., Uritboonthai W., Turner K., Siuzdak G. "A Surface Rearrangement Mechanism for Desorption/Ionization on Porous Silicon," submitted, 2006.

O'Maille G., Hoang L., Nordstrom A., Go E.P., Qin C., Siuzdak G. "Enhanced Metabolite Profiling via Chemical Derivatization and Isotope Labeling," submitted, 2006.

Nordstrom A., O'Maille G., Qin C., Siuzdak G. "Non-linear Data Alignment for UPLC-MS and HPLC-MS based Metabolomics: Quantitative Analysis of Endogenous and Exogenous Metabolites in Human Serum," *Analytical Chemistry*, 78, 7289-3295, 2006.

Go E.P., Uritboonthai W., Apon J.A., Trauger S.A., Nordstrom A., O'Maille G., Brittain S., Peters E.C., Siuzdak G. "Fluorous Affinity Tags for Selective Metabolite and Peptide Capture and Mass Detection," submitted, 2006.

Want E.J., Smith C., Siuzdak G. "Phospholipid Capture Combined with Non-Linear Chromatographic Correction for Improved Metabolite Profiling," *Metabolomics*, in press, 2006.

Fan L., Arvai A., , Cooper P.K., Iwai S., Hanaoka F., Tainer J.A. "Conserved XPB Core Structure and Motifs for DNA Unwinding: Implications for Pathway Selection of Transcription or Excision Repair," *Molecular Cell*, 22: 27-37, 2006.

Pascal J. M., Tsodikov O.V., Hura G.L., Song W., Cotner E.A., Classen S., Tomkinson A.E., Tainer J.A., Ellenberger T. "A flexible interface between DNA ligase and a heterotrimeric sliding clamp supports conformational switching and efficient ligation of DNA," *Molecular Cell*, 24:279-91, 2006.

Tsutakawa S.E., Hura G.L., Frankel K.A., Cooper P.K., Tainer J.A. "Structural analysis of flexible proteins in solution by small angle X-ray scattering combined with crystallography," *J. Structural Biology*, in press, 2006.

Chris H.Q. Ding, Xiaofeng He, and Stephen R. Holbrook, "Transitive closure and metric inequality of weighted graphs – detecting protein interaction modules using cliques," Int. J. Data Mining and Bioinformatics Vol.1, No.2. 2006.

C. Wang, C, Ding, R.F. Meraz, and S.R. Holbrook, "PSoL: a positive sample only learning algorithm for finding non-coding RNA genes," Bioinformatics, 22:2590-2596, 2006.

Chris Ding, Ya Zhang, and Stephen Holbrook, "Biclustering Protein Complex Interactions with a Biclique Finding Algorithm," 2006 IEEE International Conference on Data Mining, IEEE Computer Society Press (in press), 2006.

Chunlin Wang, Chris H.Q. Ding & Stephen R. Holbrook, "Anatomy of the Yeast Protein Interaction Network by Hierarchical Decomposition," (Submitted to Nature Biotechnology), 2006.

Ya Zhang, Chris Ding and Stephen Holbrook, "Simultaneously Relating Domains and Protein complexes through Biclique Discovery," (Submitted to Bioinformatics), 2006.

30 GTL

The MAGGIE Project: Identification and Characterization of Native Protein Complexes and Modified Proteins from *Pyrococcus furiosus*

Angeli Lal Menon[1]* (almenon@uga.edu), Farris L. Poole II,[1] Aleksandar Cvetkovic,[1] Saratchandra Shanmukh,[1] Joseph Scott,[1] Francis E. Jenney Jr.,[1] Sunia Trauger,[2,3] Ewa Kalisiak,[2,3] Gary Siuzdak,[2,3] Greg Hura,[3] John A. Tainer,[3] and **Michael W. W. Adams**[1]

[1]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, Georgia; [2]Center for Mass Spectrometry, The Scripps Research Institute, La Jolla, California; and [3]Department of Biochemistry and Molecular Biology, The Scripps Research Institute, La Jolla, California

Project Goals: Our goals are to (a) identify native multiprotein complexes (PCs) and modified proteins (MPs), such as those containing organic and/or inorganic cofactors, using native biomass of a model hyperthermophilic organism, *Pyrococcus furiosus*, by mass spectrometry in collab-

oration with Gary Suizdak, (b) to provide native samples of the more abundant PCs and MPs for characterization by small angle X-ray scattering (SAXS) in collaboration with John Tainer, (c) to use bioinformatic approaches to validate and define PCs and MPs for multiple ORF expression in collaboration with Steve Yannone, Nitin Baliga and Steve Holbrook, (d) to produce recombinant PCs on analytical and preparative scales for structural characterization in collaboration with John Tainer, (e) to design homologous and heterologous genetic approaches for the production and characterization of PCs in collaboration with Steve Yannone, Nitin Baliga (f) to design and evaluate generic protocols for PC and MP protein production in other prokaryotic systems of DOE interest, with Steve Yannone, Nitin Baliga and Steve Holbrook.

Most cellular processes are carried out by dynamic molecular machines or large protein complexes (PCs), and many of which include post-translationally-modified proteins (MPs), such as those containing organic and/or inorganic cofactors. Despite the fact that most cellular proteins exist in the form of stable or transient PCs, their composition and the ORFs that encode the components of these complexes are largely unknown. They cannot be predicted from bioinformatics analyses. In addition, no well defined techniques are currently available to unequivocally identify PCs or MPs and their individual components. Some of these issues can be resolved by determining the identity of PCs and MPs found in native proteomes. We are using the archaeon, *Pyrococcus furiosus*, a hyperthermophile that grows optimally near 100˚C, as the model organism. By analyzing the native proteome at ambient temperatures, close to 80˚C below the optimal physiological and growth temperature, the goal is to capture both stable and dynamic/transient protein complexes for identification, purification, and molecular and functional characterization.

Large scale fractionation of native *P. furiosus* biomass is being accomplished using non-denaturing, column chromatography techniques. Samples from the column fractions are being analyzed by native and denaturing PAGE, mass spectrometry (nano LC-ESI-MS/MS and MALDI-MS) and metal analyses (colorimetric and ICP-MS) to identify PCs and MPs and to determine their relative abundance in the native biomass. The more abundant PCs and MPs obtained from native biomass fractionation are being directly analyzed by Multiple Angle Light Scattering (MALS), Dynamic Light Scattering (DLS) and Small Angle X-ray Scattering (SAXS) to provide information on purity, native complex mass and subunit stoichiometry. Purified, abundant native complexes are also being used for structural characterization. The less abundant PCs and MPs and their individual components are being produced using recombinant gene expression and purification based on bioinformatic predictions and data from the native biomass analyses. The recombinant portion of the project takes advantage of the pre-existing infrastructure developed for a previous structural genomics effort with *P. furiosus*. In a preliminary pilot study almost 600 proteins were identified in fractions eluted from the first chromatographic separation of the cytoplasmic fraction from native *P. furiosus* biomass. Of these, 108 were proposed to be part of 45 potential heteromeric complexes in high abundance according to their elution behavior. A total of 29 of the 45 were previously uncharacterized, consisting of predominantly conserved hypothetical proteins, and not predicted to encode PCs. Approximately half of the fractions from the first chromatography step were subsequently fractionated by a total of 16 additional chromatography steps yielding almost 1000 distinct fractions. The nature of the PCs and MPs (particularly metal-containing proteins) that were identified and purified in this pilot study of native biomass will be described.

* Presenting author

31 GTL

The MAGGIE Project: Production and Isolation of Tagged Native/ Recombinant Multiprotein Complexes and Modified Proteins from Hyperthermophilic *Sulfolobus solfataricus*

Denise Munoz,[1] Jill Fuss,[1] Kenneth Stedman,[2] Michael W. W. Adams,[3] Gary Siuzdak,[4] Nitin S. Baliga,[5] Stephen R. Holbrook,[1] John A. Tainer,[1,6] and **Steven M. Yannone**[1]* (SMYannone@lbl.gov)

[1]Department of Molecular Biology, Lawrence Berkeley National Laboratory, Berkeley, California; [2]Center for Life in Extreme Environments, Portland State University, Portland, Oregon; [3]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, Georgia; [4]Center for Mass Spectrometry, The Scripps Research Institute, La Jolla, California; [5]Institute for Systems Biology, Seattle, Washington; and [6]Department of Biochemistry and Molecular Biology, The Scripps Research Institute, La Jolla, California

Project Goals: 1. To develop molecular biology tools to affinity tag *S. solfataricus* genes and reintroduce them into the native organism in a high-throughput manner. 2. To isolate tagged protein complexes and modified proteins from soluble and membrane fractions of *S. solfataricus* extracts. 3. To characterize protein complex components and stoichiometry by 1D/2D gel separation, mass spectrometry, and small angle X-ray scattering (SAXS).

Dynamic protein-protein interactions are fundamental to most biological processes and essential for maintaining homeostasis within all living organisms. Understanding the networks of these protein interactions is of critical importance to understanding the complexities of biological systems. The MAGGIE project was conceived, in part, as a response to the DOE GTL initiative to develop technologies to map the proteomes of model organisms. In this project we are exploiting unique characteristics of members of extremophilic Archaea to identify, isolate, and characterize multi-protein molecular machines. We have teamed expertise in mass spectrometry, systems biology, structural biology, biochemistry, and molecular biology to approach the challenges of mapping relatively simple proteomes. As part of the MAGGIE project, we are developing shuttle vectors for the extremophilic organism *Sulfolobus solfataricus* which has a growth optimum at 80˚C and pH 3.0. We are using a naturally occurring viral pathogen of this organism to engineer shuttle vectors designed for recombinant protein tagging and expression in the native *Sulfolobus* background. We are also exploiting the unique characteristics of Archaeal membranes to isolate membrane-protein complexes from native biomass. We will test the idea that the hyperthermophilic nature of *Sulfolobus* will allow U.S. to "thermally trap" protein complexes assembled at 80˚C by isolating these complexes at room temperature. Our component is interfacing with other MAGGIE components to characterize isolated proteins and protein complexes using MS/MS and small angle x-ray scattering at the advanced light source at LBNL. Ultimately, we aim to identify metabolic modules suitable to transfer specific metabolic processes between microbes to address specific DOE missions while developing generally applicable molecular and biophysical technologies for GTL.

32 GTL

Protein Complex Analysis Project (PCAP): Project Overview

Dwayne Elias,[3] Swapnil Chhabra,[1] Jil T. Geller,[1] Hoi-Ying Holman,[1] Dominique Joyner,[1] Jay Keasling,[1,2] Aindrila Mukhopadhyay,[1] Mary Singer,[1] Tamas Torok,[1] Judy Wall,[3] Terry C. Hazen,[1] Gareth Butland,[1] Ming Dong,[1] Steven C. Hall,[4] Bing K. Jap,[1] Jian Jin,[1] Susan J. Fisher,[4] Peter J. Walian,[1] H. Ewa Witkowska,[4] Lee Yang,[1] **Mark D. Biggin**[1]* (mdbiggin@lbl.gov), Manfred Auer,[1] Agustin Avila-Sakar,[1] Florian Garczarek,[1] Robert M. Glaeser,[1] Jitendra Malik,[2] Eva Nogales,[2,4] Hildur Palsdottir,[1] Jonathan P. Remis,[1] Dieter Typke,[1] Kenneth H. Downing,[1a] Steven S. Andrews,[1] Adam P. Arkin,[1,2] Steven E. Brenner,[1,2] Y. Wayne Huang,[1] Janet Jacobsen,[2] Keith Keller,[2] Ralph Santos,[1] Max Shatsky,[2] and John-Marc Chandonia[1]

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

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

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

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

* Presenting author

cal, control-theoretical, and evolutionary point of view, the role of multi-protein complexes in stress pathways involved in the biogeochemistry of soil microbes under a wide variety of conditions.

Essential to this endeavor is the development of automated high throughput methods that are robust and allow for the comprehensive analysis of many protein complexes. Biochemical purification of endogenous complexes and identification by mass spectrometry is being coupled with *in vitro* and *in vivo* EM molecular imaging methods. Because no single method can isolate all complexes, we are developing two protein purification pipelines, one the current standard Tandem Affinity Purification approach, the other a novel tagless strategy. Specific variants of each of these are being developed for water soluble and membrane proteins. Our Bioinstrumentation group is developing highly parallel micro-scale protein purification and protein sample preparation platforms, and mass spectrometry data analysis is being automated to allow the throughput required. The stoichiometries of the purified complexes are being determined and the quaternary structures of complexes larger than 250 kDa are being solved by single particle EM. We are developing EM tomography approaches to examine whole cells and sectioned, stained material to detect complexes in cells and determine their localization and structures. New image analysis methods will be applied to speed determination of quaternary structures from EM data. Once key components in the interaction network are defined, to test and validate our pathway models, mutant strains not expressing these genes will be assayed for their ability to survive and respond to stress and for their capacity for bioreduction of DOE important metals and radionuclides.

Our progress during the first year of the project includes constructing genetically altered *D. vulgaris* strains and using them to test a range of affinity tags for the purification of protein complexes and EM localization of complexes in cells, developing automated primer design algorithms for high throughput recombinant DNA and strain engineering, establishing cost effective strategies to produce up to 200 L cultures of *D. vulgaris* per week, establishing an optimized four-step tagless fractionation series for the purification of water soluble protein complexes, adopting an efficient PVDF membrane micro titer plate-based methods for mass spectrometry sample preparation, determining the structure of the 1 MDa *D. vulgaris* Pyruvate Ferredoxin Oxidoreductase complex to 17Å resolution by single particle EM, and constructing relational databases for biomass production, genetically manipulated *D. vulgaris* strains, single particle EM, and tagless complex purification. Further details on these and other results are provided in Subproject specific posters.

33 GTL

Protein Complex Analysis Project (PCAP): Multi-Protein Complex Purification and Identification by Mass Spectrometry

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

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

Project Goals: This subproject of the Protein Complex Analysis Project (PCAP) is developing several complementary high throughput pipelines to purify protein complexes from *D. vulgaris*, identify their polypeptide constituents by mass spectrometry, and determine their stoichiometries. Our goal is to determine an optimum strategy that may include elements of each purifica-

tion method. These methods will then be used as part of PCAP's effort to model stress responses relevant to the detoxification of metal and radionuclide contaminated sites.

This subproject of the Protein Complex Analysis Project (PCAP) is developing several complementary high throughput pipelines to purify protein complexes from *D. vulgaris*, identify their polypeptide constituents by mass spectrometry, and determine their stoichiometries. Our goal is to determine an optimum strategy that may include elements of each purification method. These methods will then be used as part of PCAP's effort to model stress responses relevant to the detoxification of metal and radionuclide contaminated sites.

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

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

Our results for the first year of the project are as follows.

Tagless purification of water soluble complexes. We have developed an optimized four-step fractionation scheme for the tagless purification strategy and have used it to identify and purify 15 homomeric and heteromeric water soluble protein complexes from *D. vulgaris*. We have established an efficient, highly reproducible mass spectrometry sample preparation protocol that uses 96-well PVDF multiscreen plates, which will greatly aid high throughput analysis. This sample preparation method is effective for the iTRAQ methodology we have adopted to help quantitate the relative abundances of polypeptides in different chromatographic fractions. Methods for preparing protein samples suitable for single particle EM analysis are being refined, including the use of different crosslinking reagents to stabilize complexes on EM grids. To date, five complexes have been sent to the EM Subproject for structural determination. As a result a 17 Å resolution structure of the 1 MDa Pyruvate Ferredoxin Oxidoreductase complex has been obtained.

Tagless purification of membrane complexes. We have developed an improved strategy to isolate membrane protein complexes that uses a multi-step membrane solubilization approach in which inner and outer membrane proteins are processed sequentially. The choice of detergent was shown

* Presenting author

to be critical, especially for the isolation of outer membrane protein complexes. Our methods for preparing samples for mass spectrometry analysis have been improved, particularly in the area of membrane protein native PAGE. Five membrane protein complexes have been identified and several others have been purified, ready for mass spectrometry analysis. Large-scale application of these methods is expected to facilitate the isolation and identification of substantially more complexes over the next project year.

Tandem Affinity Purification of water soluble complexes. We have commenced trials of different TAP tag combinations for protein complex purification from *D. vulgaris*. Initial tests have compared the efficiency and observed non-specific binding properties of Sequential Peptide Affinity (SPA) tag, which is composed of Calmodulin Binding Peptide (CBP) and FLAG affinity purification tags, and the Strep-TEV-FLAG (STF) tag, which is similar to SPA but with CBP replaced by a Streptavidin tag. We have confirmed that both tags can purify proteins synthesized in *D. vulgaris* and are currently testing an expanded set of tagged proteins. Once completed, high-throughput methods currently being developed will be used to construct tagged *D. vulgaris* genes rapidly and efficiently.

Automation of protein complex purification. We have developed a prototype multi-channel, native gel electrophoresis instrument for high resolution protein separation and automated band collection. This instrument can elute a protein band into a 200 µl fraction (about 60% of the band), without noticeable loss of sample. The use of this free-flow electrophoresis apparatus will greatly assist our efforts to achieve high throughput and provide an additional means of obtaining specimens in amounts appropriate for EM studies.

Mass spectrometry. Optimization of MALDI TOF/TOF MS/MS conditions is necessary to maximize the quality and consequently the information content of the data. Resolution of the precursor selection window, number of laser shots, collision energy and collision gas pressure have been evaluated from the point of view of the success rate of protein identification and quality of the iTRAQ-based quantitation. Ultimately our high throughput mass spectrometry workflow will employ highly customized information-dependent selection of precursors for MS/MS. We have begun evaluation of different aspects of iterative MS/MS acquisition routines with the aim of limiting collection of redundant data on proteins already identified and focusing on reliable quantification and identification of less abundant species. The strategy employs collection of MS and very limited MS/MS during the first iteration followed by MS/MS acquisition performed in discrete stages, with each stage building upon a combination of results of current and preceding analyses of adjacent fractions within the same protein complex separation step. We have also are developing algorithms and graphical display tools for identifying protein complexes from mass spectrometry data, including a method for cluster analysis of tagless iTRAQ data to allow for detection of comigrating polypeptides and hence putative protein complexes.

34 GTL

Protein Complex Analysis Project (PCAP): Imaging Multi-Protein Complexes by Electron Microscopy

Manfred Auer,[1] Agustin Avila-Sakar,[1] David Ball,[1] Florian Garczarek,[1] Robert M. Glaeser,[1] Jitendra Malik,[2] Eva Nogales,[1,2] Hildur Palsdottir,[1] Jonathan Remis,[1] Max Shatsky,[2] Dieter Typke,[1] and **Kenneth H. Downing**[1]* (KHDowning@lbl.gov)

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

Project Goals: The broad aim of this Subproject of PCAP is to demonstrate the feasibility of using electron microscopy for high-throughput structural characterization of multi-protein complexes in microbes of interest to DOE. One goal of this work is to characterize the degree of structural homogeneity or diversity of the multi-protein complexes purified by PCAP and to determine the spatial arrangements of individual protein components within the quaternary structure of each such complex. A second goal is to determine the spatial organization and relative locations of large multi-protein complexes within individual, intact microbes. A third goal is to determine whether whole-cell characterization by cryo-tomography can be further supplemented by electron microscopy of cell-envelope fractions and even the whole-cell contents of individual, lysed cells. Finally, plastic-section electron microscopy is used to translate as much as possible of this basic understanding to the more relevant physiological conditions, both stressed and unstressed, of planktonic and biofilm forms of microbes of interest.

The broad aim of this Subproject of PCAP is to demonstrate the feasibility of using electron microscopy for high-throughput structural characterization of multi-protein complexes in microbes of interest to DOE. Protocols and infrastructure are being developed to identify suitable candidates for structural study among the complexes isolated by the other components of the PCAP group, and to determine the spatial arrangements of individual protein components within the quaternary structure of each such complex. At a resolution of ~ 2 nm it is possible to locate the positions of individual proteins within such complexes and to then dock previously-determined atomic models of the identified proteins into the envelope of the density map. At resolutions better than 1 nm it is possible to further characterize conformational changes. We aim to increase the throughput of such structure determinations to the level that quaternary structures and docked atomic models are produced within 48 hours of purification of individual, structurally homogeneous complexes.

A second goal is to determine the spatial organization and relative locations of large multi-protein complexes within individual, intact microbes. It has quite recently been established that cryo-EM tomography can be used to produce clearly distinguishable images of larger multiprotein complexes (M_r > ~750 k) within suitably thin, intact cells. Since the cells are imaged in a nearly undisturbed condition, it is possible to count the number of such complexes in each cell as well as to characterize their spatial distribution and their association with other components of subcellular structure. Our present aim is to characterize large subcellular structures in *Desulfovibrio vulgaris* to provide a basis for understanding the morphological changes that follow various stresses.

We also employ plastic-section electron microscopy to study both planktonic and biofilm forms of microbes of interest. This approach has the advantage that it lends itself more easily to labeling – and thus localizing – genetically tagged proteins. Sectioning is also the only technique that can provide images of specimens that are too thick to image as whole-mount materials, while still retaining

* Presenting author

nanometer resolution. The ultimate goal in using plastic-section microscopy is thus to provide the most complete and accurate information possible about the status of multi-protein complexes, and to do so in a way that can then be used to improve mathematical modeling of cellular responses under the environmental conditions that require bioremediation.

In our initial work on single particle EM structures, we developed a pipeline for characterization of sample homogeneity that calls for an initial evaluation of each such specimen in uranyl acetate, in neutralized phosphotungstic acid, and in ammonium molybdate, in order minimize misleading characterizations that inevitably occur due to unwanted stain-specimen interactions (e.g. spurious aggregation). One of the more promising specimens that appeared to be suitably homogeneous and well dispersed has been fully analyzed as a demonstration of the procedures and the information provided by the analysis. The protein was identified as pyruvate-ferredoxin oxidoreductase (PFOR), which is present in *D. vulgaris* as an octamer of about 1 MDa molecular weight, while it is found as a dimer in other bacteria, including *Desulfovibrio africanus*. From EM images of protein in negative stain, a three dimensional density map was derived with resolution sufficient to unambiguously dock the x-ray crystal structure of a dimeric form of the enzyme that had previously been obtained from the *D. africanus* protein. The *D. vulgaris* amino acid sequence is found to have one insertion in a surface loop at the interface between subunits that would appear to account for the difference in oligomerization state. Further work is needed to understand the physiological significance of forming the octamer rather than dimer in *D. vulgaris*.

In order to take advantage of the genetic tools that allow tagging of specific proteins for localization by both light and electron microscopy, we are focusing on several fluorescent reagents that can be characterized in the light microscope and then photoconverted to electron dense signals for electron microscopy. This is quite a new endeavor for anaerobic bacteria such as *D. vulgaris*. Our initial focus is on morphology of biofilms in which we see a number of structures that have yet to be characterized in *D. vulgaris*. We grow biofilms of *D. vulgaris* in cellulose dialysis tubing, where the biofilms cover almost the entire interior of the tube. Samples are high pressure frozen and freeze-substituted in order to optimize preservation of structural details. Electron microscopic analysis of biofilm sections reveal loose packing of *D. vulgaris* within the biofilm EPS. Interestingly we found filamentous string-like metal precipitates near the *D. vulgaris*, which may point to structures not unlike the well-characterized *Shewanella* nanowires, which are known to be instrumental in extracellular metal reduction. Variations in the deposition patterns indicate that metal reduction activity varies between neighboring cells in biofilms. Of particular interest are strings of vesicles that appear to be extruded from the cells, similar to structures we have seen also in biofilms of *Shewanella oneidensis* and *Myxococcus xanthus*. We have developed on-grid culturing methods for fast study of such features in cell monolayers grown under various environmental conditions

We have had preliminary success in ReAsH and SNAP-labeling of several strains of *D. vulgaris* in which proteins have been tagged by members of the PCAP Microbiology group. The labeling appears promising as judged by light microscopy, and in-vitro labeling of tagged proteins after cell lysis followed by SDS PAGE suggests specific binding for the SNAP-tag reagent. First attempts at photoconversion of the fluorescence signal are currently underway.

35 GTL

Protein Complex Analysis Project (PCAP): Microbiology Subproject

Hoi-Ying Holman,[1,4] Jay Keasling,[1,2,4] Aindrila Mukhopadhyay,[1,4] Swapnil Chhabra,[1,4] Jil T. Geller,[1,4] Mary Singer,[1,4] Dominique Joyner,[1,4] Tamas Torok,[1,4] Judy Wall,[3,4] Dwayne A. Elias,[3,4] and **Terry C. Hazen**[1,4*] (tchazen@lbl.gov)

[1]Lawrence Berkeley National Laboratory, Berkeley, California; [2]University of California, Berkeley, California; [3]University of Missouri, Columbia, Missouri; and [4]Virtual Institute for Microbial Stress and Survival, Berkeley, California http://vimss.lbl.gov

Project Goals: The Microbiology Subproject of the Protein Complex Analysis Project (PCAP) provides the relevant field experience to suggest the best direction for fundamental, but DOE relevant, research as it relates to bioremediation and natural attenuation of metals and radionuclides at DOE contaminated sites. We are building on techniques and facilities established by the Environmental Stress Pathways Project (ESPP) for isolating, culturing, and characterizing Desulfovibrio vulgaris. The appropriate stressors for study will be identified and, using stress response pathway models from ESPP, the relevance and feasibility for high throughput protein complex analyses will be assessed. Two types of genetically engineered strain are being constructed: strains expressing affinity tagged proteins and knock out mutation strains that eliminate expression of a specific gene. High throughput phenotyping of these engineered strains will then be used to determine if any show phenotypic changes. Finally, we are producing large quantities of cells under different conditions and harvesting techniques for optimal protein complex analyses.

The Microbiology Subproject of PCAP provides the relevant field experience to suggest the best direction for fundamental, but DOE relevant research as it relates to bioremediation and natural attenuation of metals and radionuclides at DOE contaminated sites. This project builds on techniques and facilities established by the Virtual Institute for Microbial Stress and Survival (VIMSS) for isolating, culturing, and characterizing *Desulfovibrio vulgaris*. The appropriate stressors for study have been identified and, using stress response pathway models from VIMSS, the relevance and feasibility for high throughput protein complex analyses is being assessed. We also produce all of the genetically engineered strains for PCAP. Two types of strain are being constructed: strains expressing affinity tagged proteins and knock out mutation strains that eliminate expression of a specific gene. We anticipated producing over 300 strains expressing affinity tagged proteins every year for complex isolation and EM labeling experiments by the other Subprojects. A much smaller number of knockout mutation strains are being produced to determine the effect of eliminating expression of components of putative stress response protein complexes. Both types of engineered strains are being generated using a two-step procedure that first integrates and then cures much of the recombinant DNA from the endogenous chromosomal location of the target gene. We are developing new counter selective markers for *D. vulgaris*. This procedure will 1) allow multiple mutations to be introduced sequentially, 2) facilitate the construction of in-frame deletions, and 3) prevent polarity in operons. The Microbiology Subproject provides high throughput phenotyping of all engineered strains to determine if any show phenotypic changes. We also determine if the tagged proteins remain functional and that they do not significantly affect cell growth or behavior. The knockout mutations are tested in a comprehensive set of conditions to determine their ability to respond to stress. High throughput optimization of culturing and harvesting of wild type cells and all engineered strains are used to determine the optimal time points, best culture techniques, and best techniques for harvesting cultures using real-time analyses with synchrotron FTIR spectromicroscopy, and other methods. Finally, we are producing large quantities of cells under different conditions and harvesting techniques for opti-

* Presenting author

mal protein complex analyses. To insure the quality and reproducibility of all the biomass for protein complex analyses we use extreme levels of QA/QC on all biomass production. We expect to do as many as 10,000 growth curves and 300 phenotype microarrays annually and be producing biomass for 500-1000 strains per year by end of the project. Each biomass production for each strain and each environmental condition will require anywhere from 0.1 – 400 L of culture, and we expect more than 4,000 liters of culture will be prepared and harvested every year. The Microbiology Subproject is optimizing phenotyping and biomass production to enable the other Subprojects to complete the protein complex analyses at the highest throughput possible. Once the role of protein complexes has been established in the stress response pathway, we will verify the effect that the stress response has on reduction of metals and radionuclides relevant to DOE.

During the last year, the Microbiology Subproject has supplied more than 30 sets of *D. vulgaris* cultures on biofilms for EM analysis, more than twenty 1-5 L cultures of biomass for water-soluble protein complex purification studies, and two 100 L and one 200 L cultures for membrane protein complex purification. We have designed and implemented a continuous culture system that enables U.S. to produce more than 300 L of anaerobic mid log phase *D. vulgaris* in as little as 5 days, including harvesting and QA techniques that maxmize reproducibility of all biomass produced. The goal of incorporating different affinity or tandem affinity (TAP) protein tags into three genes to determine the best tag for use in the PCAP project will be complete by the end of 2006. These include the Strep-tag® (IBA) for streptavidin-binding, the SPA-tag (a.k.a. CTF) that consists of a calmodulin binding motif, tobacco etch virus protease (TEV) and 3X FLAG affinity, as well as a combination of these that replaces the calmodulin binding with the Strep-tag® resulting in STF. The three genes to test are the dissimilatory sulfite reductase subunit C, pyruvate ferrodoxin oxidoreductase subunit B and ATP synthase subunit C. Additionally, several other gene targets have been identified through close collaboration with the VIMSS/ESPP group at LBNL and are currently being tagged. To determine localization of a given gene product in the cell, we have utilized both the tetracysteine and SNAP-tag™ (Covalys) in cooperation with the EM group of the PCAP project. Currently the total number of genes tagged with CTF are 3, with STF 6, with strep 20, with tetracysteine 8 and with SNAP 3. We are currently attempting to construct an ordered library for tagging in DvH. By doing so, this will allow for a relatively small number of *E. coli* clones to carry all of the genes for DvH, thereby reducing the overall workload and paving the way for higher throughput tagging. We also compared several cloning strategies for producing tagged constructs for the plasmid insertion strategy in *D. vulgaris*. The two-step TOPO-GATEWAY strategy (Invitrogen) was identified as the most economical commercially available conventional-cloning strategy amongst these. We also developed a workflow for high throughput production of tagged strains of *D. vulgaris*. This workflow was based on the TOPO-GATEWAY strategy in combination with current technology for transformation of *D. vulgaris*. We constructed custom destination vectors carrying the tags SPA and STF (to realize the GATEWAY step)—these are not available commercially. We tested the workflow through all the steps for a set of 10 randomly chosen genes from the *D. vulgaris* genome. Based on generated sequences, five of these were successfully tagged with the SPA tag. We also tested a commercially available 96-well electroporation device (BTX) for high throughput transformation of *D. vulgaris*. This system was found unsuitable. We are currently developing a custom solution for this. We also collaborated with the Subgroup D (Computational Core) for the development of automated algorithms for: 1) Primer identifications based on gene locations within operons for PCR amplifications in 96-well format, and 2) Sequence alignments for identifying errors in the amplifications or cloning steps in the workflow. All of the tagged strains constructed this year have been characterized using phenotypic microarrays (PM), and the *D. vulgaris* megaplasmid minus strain (MP(-)) being used in the electroporation studies is being aggressively characterized for all differences including stress responses by the ESPP project.

36 GTL

Protein Complex Analysis Project (PCAP): High Throughput Strategies for Tagged-Strain Generation in *Desulfovibrio vulgaris*

Swapnil Chhabra[1]* (SRChhabra@lbl.gov), Gareth Butland,[1] Dwayne Elias,[2] Aindrila Mukhopadhyay,[1] John-Marc Chandonia,[1] Jay Keasling,[1,3] and **Judy Wall**[2]

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

Project Goals: As part of the microbiology core of the Protein Complex Analysis Project (PCAP) our goal is to develop a technological platform for creating a library of *D. vulgaris* mutant strains expressing tagged proteins at high throughput. Based on the workflow designed around the TOPO-GATEWAY strategy, we will produce a hundred constructs carrying the STF tag which will be transformed in *D. vulgaris* to create a tagged strain library. We are also exploring an alternative high throughput strategy using an ordered library of *D. vulgaris*.

Most cellular processes are mediated by a host of different proteins interacting with each other in the form of complexes. As a follow-up to the functional genomics analyses of stress response pathways in *Desulfovibrio vulgaris*, are studying the role of protein complexes in this sulfate reducing bacterium which has been found to exist in several DOE waste sites. As part of the microbiology core of the Protein Complex Analysis Project (PCAP) our current goal is to develop a technological platform for creating a library of *D. vulgaris* mutant strains expressing tagged proteins. We are currently exploring two approaches to achieve this goal. The first strategy involves the use of plasmid constructs carrying single target genes using the two-step TOPO-GATEWAY cloning approach (Invitrogen). The first step in the strategy involves generation of an entry vector carrying the gene of interest (GOI) via TOPO cloning. The second step involves transfer of the GOI from the entry vector to a suitable destination vector (carrying the tag of choice) through an in-vitro recombination reaction. The second strategy involves the use of an ordered library of *D. vulgaris* modified using a lambda-red phage system. Library constructs are modified, in a strain of *E. coli* expressing the lambda-red recombination system, using linear PCR products specifically engineered to recombine into the 3´ end of the gene of interest. These PCR products, when inserted into the gene of interest modify the coding sequence of the gene to encode a C-terminal fusion protein bearing the tag of choice. Constructs from both strategies are transferred to *D. vulgaris* via electroporation and replacement of the wild type copy of the gene of interest with the tagged version is selected. These approaches would enable U.S. to rapidly and efficiently modify the *D. vulgaris* genome and express tagged proteins at their native levels.

* Presenting author

37 GTL

Protein Interaction Reporters and Outer Membrane Cytochrome C

James E. Bruce[1]* (james_bruce@wsu.edu), Haizhen Zhang,[1] Natalia Zakharova,[1] Xiaoting Tang,[1] Gerhard R. Muske,[1] Liang Shi,[2] James K. Fredrickson,[2] Nikola Tolic,[2] and Gordon A. Anderson[2]

[1]Department of Chemistry, Washington State University, Pullman, Washington http://www.wsu.edu/proteomics and [2]Environmental Molecular Science Laboratory, Pacific Northwest National Laboratory, Richland, Washington http://www.emsl.pnl.gov

Project Goals: The overall aim of this project is to develop the PIR approach to identify protein interactions in cells and apply this new methodology to the study of systems relevant to the DOE mission.

We have developed a unique chemical cross-linking system that employs novel compounds that we call "Protein Interaction Reporters" or PIRs that can help identify interactions among proteins in complex biological systems. The incorporation of mass spectrometry-cleavable bonds in the PIR structure allows release of labeled, intact peptides that can subsequently be analyzed and identified with high mass measurement accuracy or tandem MS methods. This approach enables a new method of analysis of cross-linked proteins from complex mixtures since protein identification is established via released peptide measurements, while interaction information is established through PIR-labeled peptide analysis. Since interactions among proteins in cells are critical determinants of overall function, the ability to identify and measure protein interactions in a cellular system with PIR technology can significantly improve molecular-level comprehension of biological function. The overall aim of this project is to develop the PIR approach to identify protein interactions in cells and apply this new methodology to the study of systems relevant to the DOE mission. One area where this approach will have positive impact is in the study of electron transfer mechanisms in bacteria such as *Shewanella oneidensis*. For example, a central question relevant to many biological systems with potential for bioremediation or bio-energy production revolves around the protein interaction pathways that facilitate novel mechanisms of electron transport. Improved understanding of these pathways can conceivably guide bioengineering efforts to result in improved electron transport properties.

We have previously applied the PIR strategy to map interactions in a model noncovalent complex and illustrate fesibility[1]. More recently, we have applied PIR technology to the microbial system, *Shewanella oneidensis MR-1*, and identified more than 380 proteins that are labeled during on-cell PIR reactions[2]. We have also developed additional PIR structures that provide increased chemical diversity to further improve the number and type of proteins that can be studied with the PIR approach[3]. This report will present these recent developments and the application of the PIR strategy with efforts focused on the identification of interactions of proteins known to be critical to electron transport, such as outer membrane cytochrome (OMC) proteins. This report will also will highlight our efforts to study OMC proteins and their interactions both *in vitro* and in cells using the PIR strategy combined with immunoaffinity methods.

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

References

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

2. Tang, X. Yi, W., Munske, G.R. Adhikari, D.P. Zakharova, N.L. Bruce, J.E. Profiling the membrane proteome of *Shewanella oneidensis* MR-1 with new affinity labeling probes. *J. Proteome Research*, in press.
3. Chowdhury, S.M. Munske, G.R. Tang, X. Bruce, J.E. Collisionally Activated Dissociation and Electron Capture Dissociation of Several Mass Spectrometry-Identifiable Chemical Cross-Linkers. *Anal. Chem.* in press.

38 GTL

Technologies for Comprehensive Protein Production

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

Argonne National Laboratory, Argonne, Illinois

Progress in genome sequencing has accentuated the importance of high throughput proteomic strategies for identification of cellular function. Many proteomic technologies such as functional screens, structure determination or interaction mapping use purified proteins as a starting point. Although, recent advances in expression technology have significantly increased our capability for the expression of microbial proteins, a significant fraction of the most interesting proteins encoded by the genome still cannot be expressed in an experimentally usable form. We have developed technologies to optimize the expression of proteins and protein domains from prokaryotic and eukaryotic organisms. Purity is will be assessed by SDS-PAGE stained with Coomassie Brilliant Blue. These proteins are a valuable resource for characterization studies and for structural and functional studies.

One application of this resource is the development of new approaches for in vitro production, validation, and characterization of protein complexes. Identification and characterization of protein complexes is an important component of several GTL science programs. The complexity of bacterial proteome and our limited knowledge of the number and nature of interacting proteins indicate multiple strategies must be applied to obtain a complete understanding of the number and function of cellular interacting proteins. In a preliminary study, we identified a preliminary set of 50 putative interacting proteins in *S. oneidensis* based on homology to known interacting proteins from other bacteria. These putative interacting protein pairs were validated for complex formation using an in vitro interaction assay in a manual format and via high density protein arrays for identifying interactions on protein biochips. The results indicate that many of the interactions found in *E. coli* could be rapidly validated for other organism using a high throughput approach. This set of putative interacting proteins is also being used to development of more efficient approaches for co-expression of proteins. Current expression systems sometimes produce protein subunits or interacting proteins in an insoluble form. We intend use interaction pairs from the primary screen with an insoluble or low solubility partner will be screened for improvement in production of soluble protein complexes using two co-expression strategies.

* Presenting author

39 GTL

Next-Generation Cell-Permeable Multiuse Affinity Probes (MAPs) and Cognate Tags: Applications to Bioenergy and Metabolic Engineering

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

Pacific Northwest National Laboratory, Richland, Washington

Project Goals: Our long-term goal is to develop the necessary reagents and technology for the rapid identification of a substantial fraction of the multiprotein complexes in an organism, thereby permitting a systems level understanding of signaling complexes that allow microorganisms to adapt to diverse environmental conditions. Critical to this goal is the development of new affinity reagents and protein tags that permit the rapid identification and validation of protein complexes. To accomplish these goals, we propose to develop protein tags and associated multiuse affinity probes, which promise both the ability to isolate and characterize protein complexes as well as the examination of protein-protein interactions in living microbes. We propose to extend the design of affinity reagents to include cell permeable and fluorescent photo-crosslinking reagents that permit efficient stabilization and visualization of bacterial protein complexes. Crosslinking reagents will be designed that i) react with known chemistries targeting specific protein side-chains with high yield and ii) provide cleavage sites to facilitate mass spectrometric identification of crosslinked peptides. Initial efforts will focus these technologies to identify protein complexes in *S. oneidensis* MR-1, an organism of considerable interest to DOE and the subject of multiple projects in the Genomics:GTL program (http://genomicsgtl.energy.gov/). Utilization of novel affinity reagents that become fluorescent upon binding to engineered tags will permit quantitation of expressed proteins and purification and stabilization of protein complexes. Ultimately high-throughput data collection of time-dependent changes in protein complex formation in response to environmental conditions will be available, thus permitting a systems level understanding of how microbes adapt to environmental change.

Protein-protein interactions are the foundation of the metabolic and regulatory pathways in all organisms. Before organisms can be harnessed in bioremediation and bioenergy uses, robust and scalable methodologies are needed that permit the facile identification of these pathways. To this purpose, we have developed a multi-use affinity probe (MAP) technology platform in which engineering a single encoded peptide tag onto a protein permits the identification and validation of protein complexes *in vitro* and *in vivo* in ways amenable to high-throughput scale-up through the addition of robotics. This platform overcomes prior limitations regarding the need to reengineer and perturb biomolecular systems for proteomic and imaging applications. In previous years, we have matured our technology for (i) protein complex isolation and (ii) *in vitro* validation, as well as showing how photoactivatable crosslinking moieties appended to MAPs provide a means to stabilize low-affinity binding proteins prior to cell lysis, providing the first robust means to mediate *in-vivo* cross-linking. In addition, following covalent modification fluorophore-assisted light inactivation (FALI) permits the selective knock-out of protein activities to aid in both the identification of their functions as well as the modulation of metabolic networks (*1*).

During the last year, we have extended the toolkit by adding new, bright photostable fluorescent probes based on a cyanine scaffold which are targeted to unique peptide tags and can thus be used simultaneously with first generation probes for parallel processing of tagged proteins to i) isolate and

identify the composition and size of multiple protein complexes and ii) image protein locations and associations in cells using multicolor and single molecule measurements (*2-6*). Further, the MAP toolkit was used to discover new aspects of bacterial transcription regulatory machinery and to isolate and identify functions of outer membrane cytochromes associated with electron transfer to iron oxides (e.g., hematite) from *Shewanella oneidensis*, which has important implications with respect to both bioremediation and bioenergy (*4, 7-10*).

MAPs permit the isolation of intact protein complexes, whose release from affinity matrices by mild reducing conditions allows complementary structural and functional measurements, while retaining weak binding interactions (*9, 10*). Among the important applications of this technology, extensive binding interactions between protein machinery (e.g., RNA polymerase) and regulatory proteins have been identified that are altered in response to environmental changes under controlled growth conditions in chemostats, and have important implications with respect to metabolic engineering principles (*9, 10*). We find that RNA polymerase from *Shewanella oneidensis* exists as a large supramolecular complex with an apparent mass in excess of 1.4 MDa, whose protein composition substantially changes in response to growth conditions. In comparison to suboxic conditions, a larger number of binding partners associate with RNA polymerase under aerobic conditions, where cellular growth rates are limited by the rates of ribosome biosynthesis (*11, 12*). In addition to known regulators of RNA polymerase function, binding partners include a surprisingly large number of metabolic enzymes associated with ATP synthesis, nucleotide metabolism, and the biosynthesis of stable RNA (i.e., tRNA and rRNA). Our identification of cytosolic subunits of membrane proteins with the RNA polymerase complex is consistent with recent structural data demonstrating the membrane localization of RNA polymerase in rapidly growing cells (*13*). In contrast, under suboxic conditions we observe a reduced number of protein associations, which is consistent with the observed disruption of RNA polymerase in condensed structures near the membrane under suboptimal growth conditions (*13*). In conclusion, these measurements demonstrate an unexpected functional linkage between RNA polymerase and enzymes associated with tRNA processing, nucleotide metabolism, and energy biosynthesis, which we propose to be necessary for optimal transcriptional rates and dependent on growth conditions.

Likewise, we have used this technology to isolate and identify the functions of outer member cytochromes OmcA and MtrC associated with electron transfer to iron oxides (e.g., hematite) from (*4, 7, 8*). Complementary structural and functional measurements are possible through the ability of the identified tags to promote selective binding to metal surfaces (*14*). OmcA was shown to directly associate with added hematite, as evidenced by the co-sedimentation of approximately 40% of the OmcA in solution with hematite particles upon centrifugation, permitting a determination of the binding affinity through the measurement of the concentrations of OmcA bound to hematite relative to that in water (i.e., the partition coefficient). We find that approximately 0.2 mg OmcA binds per mg of hematite (i.e., 2.5 nmol OmcA per mg hematite). In association with hematite, OmcA is catalytically active: oxidation of protein hemes, as measured from time-dependent changes in the α-Soret absorption peak at 552 nm, directly tracks with protein binding to hematite under anoxic conditions with a maximal activity of about 60 nmol mg^{-1} OmcA min^{-1}. Since OmcA can be directly reduced by NADH and other metabolic cofactors (*7*), the high-affinity interaction between OmcA and hematite provides a means to couple the generation of reducing power to an electrode surface. In summary, we have shown that purified OmcA binds and densely covers the surface of hematite, and reduces Fe(III) with a maximal velocity of approximately 60 nmol / mg min, which corresponds to an electron flux of about 10^{13} electrons /cm^2/s that approaches observed fluxes in the most efficient bioreactors (*15, 16*).

* Presenting author

References

1. Yan, P., Xiong, Y., Chen, B., Negash, S., Squier, T. C., and Mayer, M. U. (2006) Fluorophore-assisted light inactivation of calmodulin involves singlet-oxygen mediated cross-linking and methionine oxidation. *Biochemistry 45*, 4736-48.
2. Cao, H., Xiong, Y., Wang, T., Chen, B., Squier, T. C., and Mayer, M. U. (2007) A Cy3-based biarsenical fluorescent probe with a unique peptide binding motif., *J. Am. Chem. Soc., submitted.*
3. Stenoien, D., Knyushko, T., Londono, M., Opresko, L., Mayer, M. U., Squier, T. C., and Bigelow, D. J. (2007) Cellular trafficking of phospholamban and formation of functional sarcoplasmic reticulum during myocyte differentiation. *Am. J. Physiol., submitted.*
4. Xiong, Y., Shi, L., Chen, B., Mayer, M. U., Lower, B. H., Londer, Y., Bose, S., Hochella, M. F., Fredrickson, J. K., and Squier, T. C. (2006) High-affinity binding and direct electron transfer to solid metals by the *Shewanella oneidensis* MR-1 outer membrane c-type cytochrome OmcA. *J Am Chem Soc 128*, 13978-9.
5. Chen, B., Mayer, M. U., Cao, H., Yan, P., Mahaney, J. E., and Squier, T. C. (2007) Selective labeling of cytosolic and membrane proteins using cell-permeable biarsenical probes ReAsH-EDT2 and FlAsH-EDT2 *Anal. Biochem., submitted.*
6. Chen, B., Mayer, M. U., and Squier, T. C. (2007) Identification of an orthogonal peptide binding motif for biarsenical dyes. . *Bioconjugate Chemistry, submitted.*
7. Shi, L., Chen, B., Wang, Z., Elias, D. A., Mayer, M. U., Gorby, Y. A., Ni, S., Lower, B. H., Kennedy, D. W., Wunschel, D. S., Mottaz, H. M., Marshall, M. J., Hill, E. A., Beliaev, A. S., Zachara, J. M., Fredrickson, J. K., and Squier, T. C. (2006) Isolation of a high-affinity functional protein complex between OmcA and MtrC: Two outer membrane decaheme c-type cytochromes of *Shewanella oneidensis* MR-1. *J Bacteriol 188*, 4705-14.
8. Shi, L., Lin, J. T., Markillie, L. M., Squier, T. C., and Hooker, B. S. (2005) Overexpression of multiheme C-type cytochromes. *Biotechniques 38*, 297-9.
9. Mayer, M. U., Shi, L., and Squier, T. C. (2005) One-step, non-denaturing isolation of an RNA polymerase enzyme complex using an improved multi-use affinity probe resin. *Mol Biosyst 1*, 53-6.
10. Verma, S., Xiong, Y., Mayer, M. U., and Squier, T. C. (2007) Remodeling of bacterial RNA polymerase complex in response to environmental conditions *Biochemistry, submitted.*
11. Cabrera, J. E., and Jin, D. J. (2003) The distribution of RNA polymerase in *Escherichia coli* is dynamic and sensitive to environmental cues. *Mol Microbiol 50*, 1493-505.
12. Liu, M., Durfee, T., Cabrera, J. E., Zhao, K., Jin, D. J., and Blattner, F. R. (2005) Global transcriptional programs reveal a carbon source foraging strategy by *Escherichia coli*. *J Biol Chem 280*, 15921-7.
13. Jin, D. J., and Cabrera, J. E. (2006) Coupling the distribution of RNA polymerase to global gene regulation and the dynamic structure of the bacterial nucleoid in *Escherichia coli*. *J Struct Biol*, doi:10.1016/j.jsb.2006.07.005.
14. Wigginton, N. S., Rosso, K. M., Lower, B. H., Shi, L., and Hochella, M. F. (2007) Electron tunneling properties of outer membrane decaheme cytochromes from *Shewanella oneidensis*., *Geochimica et Cosmochimica Acta, in press.*
15. Ringeisen, B. R., Henderson, E., Wu, P. K., Pietron, J., Ray, R., Little, B., Biffinger, J. C., and Jones-Meehan, J. M. (2006) High power density from a miniature microbial fuel cell using *Shewanella oneidensis* DSP10. *Environ Sci Technol 40*, 2629-34.
16. Viamajala, S., Peyton, B. M., Apel, W. A., and Petersen, J. N. (2002) Chromate/nitrite interactions in *Shewanella oneidensis* MR-1: evidence for multiple hexavalent chromium [Cr(VI)] reduction mechanisms dependent on physiological growth conditions. *Biotechnol Bioeng 78*, 770-8.

40 GTL

Advanced Data Analysis Pipeline for Determination of Protein Complexes and Interaction Networks at the Genomics:GTL Center for Molecular and Cellular Systems

Kevin K. Anderson,[2] Deanna L. Auberry,[2] William R. Cannon[2]* (William.Cannon@pnl.gov), Don S. Daly,[2] Brian S. Hooker,[2] Gregory B. Hurst,[1] Jason E. McDermott,[2] W. Hayes McDonald*[1] (McDdonaldWH@ornl.gov), Dale A. Pelletier,[1] Denise D. Schmoyer[1]* (SchmoyerDD@ornl.gov), Julia L. Sharp,[3] Mudita Singhal[2]* (Mudita.Singhal@pnl.gov), Ronald C.Taylor[2]* (Ronald.Taylor@pnl.gov), **Michelle V. Buchanan**[1] (BuchananMV@ornl.gov)

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [2]Pacific Northwest National Laboratory, Richland, Washington; and [3]Montana State University, Bozeman, Montana

Project Goals: The Center for Molecular and Cellular Systems (CMCS) focuses primarily on the objectives outlined in Goal 1 of the DOE Genomics:GTL program. The core of the CMCS is a high throughput pipeline for the identification of protein complexes, currently focusing on *Rhodopseudomonas palustris*. The pipeline employs affinity isolation coupled with mass spectrometry to identify protein interactions. Computational tools are used to assess the significance of identified interactions. A dynamic research program supports the goals of the CMCS by focusing on the development and implementation of improved capabilities for complex isolation, molecular level identification of the complexes, and critical bioinformatics and computing capabilities. These efforts are focused on constructing a knowledge base that can provide insight into the relationship between the complement of protein complexes in these microbes and their biological function.

The Genomics:GTL Center for Molecular and Cellular Systems (CMCS) is a DOE Center whose mission is to determine protein complexes and interaction networks from microbial systems. Currently the center is focusing on protein interactions involved in nitrogen fixation and metabolism in *Rhodopseudomonas palustris*. The center uses an affinity purification approach (refer to poster **Global survey of protein-protein interactions in *Rhodopseudomonas palustris***) to identify protein interactions in a robust, high-throughput manner. In this process, bait proteins along with co-purifying prey proteins are extracted from the cellular milieu. The resulting protein mixture is analyzed by HPLC coupled with tandem mass spectrometry.

The pipeline for data analysis begins with a laboratory information management system (LIMS) to capture links to the MS/MS data, peptides/proteins identified from those data, and descriptions regarding biological and assay conditions (metadata). The LIMS is the central data repository for all information related to processing and analysis of CMCS samples. It maintains a detailed history for each sample by capturing processing parameters, protocols, stocks, QA/QC tests and analytical results for the complete life cycle of the sample. Project and study data are also maintained to define each sample in the context of the research tasks it supports.

The resulting lists of potentially interacting prey proteins identified from MS/MS are statistically analyzed within a software environment (*Sebini*) specifically designed for working with biological networks. The prey protein lists are cross-tabulated by bait protein to form a prey-by-bait frequency matrix. The frequency pattern across a given row (prey) shows the associations between a prey and the baits. Interpretation of this pattern depends on the selected baits. Pattern uniformity is tested with a binomial-based likelihood-ratio test. Test significance is assessed by Monte Carlo simulation

* Presenting author

where the false discovery rate is controlled. Prey protein candidates are assigned to "specific" and "non-specific" classes based on the likelihood-ratio test. Bayes estimates of the confidence of the inferred associations are estimated for each bait-prey pair. Modeling assumptions are investigated and conservative parameter estimates are made using Monte Carlo simulations.

The resulting protein networks are captured in a database within the software environment where information on the nodes (proteins) and edges (interactions) is linked to external and internal bioinformatics data, such as information on interologs derived from the *Bioverse* system, which provides additional information on protein interactions derived from orthologous proteins in other model systems. The joint analysis of experimental data and multiple sources of bioinformatically-derived information is accomplished through collective analysis of biological interaction networks (*Cabin*), a plug-in for the Cytoscape program. Protein interaction networks along with relevant data captured at multiple stages of the data analysis pipeline will be available for download at the project website (Refer to poster **The Microbial Protein-Protein Interaction Database-MiPPI**). A demonstration will be held at the workshop.

41 GTL

Global Survey of Protein-Protein Interactions in *Rhodopseudomonas palustris*

Dale A. Pelletier[1]* (pelletierda@ornl.gov), Gregory B. Hurst,[1] Linda J. Foote,[1] Trish K. Lankford,[1] Catherine K. McKeown,[1] Tse-Yuan S. Lu,[1] Elizabeth T. Owens,[1] Denise D. Schmoyer,[1] Manesh B. Shah,[1] Jennifer L. Morrell-Falvey,[1] Brian S. Hooker,[2] Stephen J. Kennel,[1] W. Hayes McDonald,[1] Mitchel J. Doktycz,[1] Deanna L. Auberry,[2] William R. Cannon,[2] Kenneth J. Auberry,[2] H. Steven Wiley,[2] and **Michelle V. Buchanan**[1]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee and [2]Pacific Northwest National Laboratory, Richland, Washington

Project Goals: The Center for Molecular and Cellular Systems (CMCS) focuses primarily on the objectives outlined in Goal 1 of the DOE Genomics:GTL program. The core of the CMCS is a high throughput pipeline for the identification of protein complexes, currently focusing on *Rhodopseudomonas palustris*. The pipeline employs affinity isolation coupled with mass spectrometry to identify protein interactions. Computational tools are used to assess the significance of identified interactions. A dynamic research program supports the goals of the CMCS by focusing on the development and implementation of improved capabilities for complex isolation, molecular level identification of the complexes, and critical bioinformatics and computing capabilities. These efforts are focused on constructing a knowledge base that can provide insight into the relationship between the complement of protein complexes in these microbes and their biological function.

The goal of the Center for Molecular and Cellular Systems (CMCS) is to identify protein-protein interaction networks, which form the molecular basis of biological function, in environmentally relevant bacterial species in support of the Genomics:GTL program. *Rhodopseudomonas palustris* is a metabolically diverse anoxygenic phototrophic bacterium that is emerging as a model system for nitrogenase-mediated hydrogen production. This process requires several metabolic and regulatory networks to be integrated within the cell, including nitrogen metabolism, photosynthesis and carbon metabolism. To better understand the interactions among these processes, we have begun mapping

protein-protein interactions of photoheterotrophically grown *R. palustris.* Toward this goal, we have developed and implemented a methodology for systematically identifying the proteins that interact with an affinity-tagged "bait" protein expressed from a plasmid introduced into *R. palustris*. The steps in this methodology include target or "bait" selection, primer design, PCR amplification, cloning, transformation, batch culture, lysis, affinity isolation, protein identification, and statistical filtering. Here we will present results on interactions identified by this approach.

To date, we have successfully cloned ~1000 *R. palustris* open reading frames, purified over 250 different affinity-tagged gene products, and identified their protein interaction partners from cultures of *R. palustris* grown under anaerobic photoheterotrophic growth conditions, in the presence or absence of fixed nitrogen. Interactors identified by this approach include homologues of a number of well-characterized protein complexes involved in known metabolic networks, such as nitrogen metabolism (Mo-nitrogenase, FixABCX with a possible role in electron transfer to nitrogenase, GlnK2-AmtB2, GlnK2-GlnB), carbon metabolism (2-oxoglutarate dehydrogenase, succinate dehydrogenase, succinyl-CoA synthetase, tryptophan synthase), transcription (DNA-directed RNA polymerase), chaperones (GroES-EL, DnaK-GrpE), and energy generation (F1F0 ATPase, subunits of NADH dehydrogenase). Novel putative interactions were also identified, including interactions among four anaerobically induced proteins encoded by RPA2334, RPA2335, RPA2336 and RPA2338; interactions between a conserved unknown RPA3193 and a putative acetyltransferase RPA3194; and interactions among conserved unknowns RPA1244, RPA1243 and RPA1246.

We are applying other approaches, including experimental, literature-based and bioinformatics predictions to verify these high-throughput interaction data (refer to poster **Advanced Data Analysis Pipeline for Determination of Protein Complexes and Interaction Networks at the Genomics: GTL Center for Molecular and Cellular Systems)**. A web resource for these data will be publicly available in February 2007 (refer to poster **The Microbial Protein-Protein Interaction Database-MiPPI**; www.ornl.gov/sci/GenomestoLife/). These results demonstrate the utility of data emerging from the CMCS for confirming known interactions, as well as for generating hypotheses about potentially novel protein-protein interactions. The identification of protein interactions will aid in elucidation of biological interaction networks and possibly in predicting protein function.

42 GTL

The Microbial Protein-Protein Interaction Database (MiPPI)

Denise D. Schmoyer[1]* (schmoyerdd@ornl.gov), Sheryl A. Martin,[1] Gregory B. Hurst,[1] Manesh B. Shah,[1] Dale A. Pelletier,[1] W. Hayes McDonald,[1] William R. Cannon,[2] Deanna L. Auberry,[2] and **Michelle V. Buchanan**[1]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee and [2]Pacific Northwest National Laboratory, Richland, Washington

Project Goals: The Center for Molecular and Cellular Systems (CMCS) focuses primarily on the objectives outlined in Goal 1 of the DOE Genomics:GTL program. The core of the CMCS is a high throughput pipeline for the identification of protein complexes, currently focusing on *Rhodopseudomonas palustris*. The pipeline employs affinity isolation coupled with mass spectrometry to identify protein interactions. Computational tools are used to assess the significance of identified interactions. A dynamic research program supports the goals of the CMCS by focusing on the development and implementation of improved capabilities for complex isolation, molecular

* Presenting author

level identification of the complexes, and critical bioinformatics and computing capabilities. These efforts are focused on constructing a knowledge base that can provide insight into the relationship between the complement of protein complexes in these microbes and their biological function.

The Microbial Protein-Protein Interaction Database (MiPPI) is a publicly accessible database of microbial protein-protein interactions experimentally detected at the Genomics:GTL Center for Molecular and Cellular Systems (CMCS). The primary experimental method used at the CMCS is affinity-based isolation combined with mass spectrometry (refer to poster **Global Survey of Protein-Protein Interactions in *Rhodopseudomonas palustris***). As of December 2006 we have performed over 500 endogenous affinity-tagged experiments which represent over 300 different bait proteins in *Rhodopseudomonas palustris* and *Shewanella oneidensis*. Our goal is to provide the highest quality protein interaction data to the biological community for the identification of cellular networks and ultimately biological function.

MiPPI stores the results of mass spectrometric protein identifications as DTASelect output files, as well as statistical evaluation of protein-protein interactions (refer to poster **Advanced Data Analysis Pipeline for Determinationof Protein Complexes and Interaction Networks at the Genomics: GTL Center for Molecular and Cellular Systems**). MiPPI is linked to the CMCS laboratory information management system (LIMS) which maintains sample metadata, from cloning through MS analysis. The first public data release is scheduled for February 2007 and includes all center results collected through November 2006. This release contains biological and technical replicates of more than 300 bait proteins collected from over 500 pulldown experiments and over 900 mass spectrometric analyses. The database includes over 50,000 observed protein-protein interactions. Updates to MiPPI will be released semiannually.

Beginning in February 2007, the web interface (www.ornl.gov/sci/GenomestoLife/) to the MiPPI database will provide online searches by protein or protein-protein interaction, and will include a protein-protein interaction viewer for the observed interactions. Mass spectrometry results and corresponding metadata will be provided for download in mzXML and DTASelect file formats. Identified protein-protein interactions including the statistical analysis scores will be provided for download in delimited text file format.

43 GTL

Advances in Coverage and Quality for High-Throughput Protein-Protein Interaction Measurements

Jennifer Morrell-Falvey,[1] Mitchel J. Doktycz,[1] Dale A. Pelletier,[1] Linda J. Foote,[1] Elizabeth T. Owens,[1] Sankar Venkatraman,[1] W. Hayes McDonald[1]* (mcdonaldwh@ornl.gov), Brian S. Hooker,[2] Chiann-Tso Lin,[2] Kristin D. Victry,[2] Deanna L. Auberry,[2] Eric A. Livesay,[2] Daniel J. Orton,[2] H. Steven Wiley,[2] and **Michelle V. Buchanan**[1]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee and [2]Pacific Northwest National Laboratory, Richland, Washington

Project Goals: The Center for Molecular and Cellular Systems (CMCS) focuses primarily on the objectives outlined in Goal 1 of the DOE Genomics:GTL program. The core of the CMCS is a high throughput pipeline for the identification of protein complexes, currently focusing on *Rho-*

***dopseudomonas palustris*. The pipeline employs affinity isolation coupled with mass spectrometry to identify protein interactions. Computational tools are used to assess the significance of identified interactions. A dynamic research program supports the goals of the CMCS by focusing on the development and implementation of improved capabilities for complex isolation, molecular level identification of the complexes, and critical bioinformatics and computing capabilities. These efforts are focused on constructing a knowledge base that can provide insight into the relationship between the complement of protein complexes in these microbes and their biological function.**

The overarching goal of the Center for Molecular and Cellular Systems (CMCS) is to identify protein interaction networks that form the molecular basis of biological function in microbes. To accomplish this goal, we have established a high-throughput analysis pipeline that is centered on generalized affinity-based isolation of protein interactors combined with mass spectrometric identification of the interacting protein components (refer to poster **Global survey of protein-protein interactions in *Rhodopseudomonas palustris***). While this approach has proven successful for identifying a large number of protein interactions within the photoheterotrophic bacterium *Rhodopseudomonas palustris*, interactions among some classes of proteins (e.g. membrane-associated or low abundance) remain difficult to detect. Approaches for overcoming these technological challenges as well as increasing throughput are needed for accomplishing our goal of identifying protein interaction networks with high confidence. Here we describe several strategies that we are developing to address these issues.

Our pipeline currently employs a tandem affinity tag comprised of 6XHis and the V5 epitope. Improvements in the affinity and specificity of the affinity tags used for labeling proteins will impact throughput by allowing the use of smaller culture volumes and improving detection of lower abundance proteins. In addition to improved affinity, the tags should be compatible with elution of the bound complex under nondenaturing conditions and amenable to automation using 96-well based robotic handling and microfluidic manipulations. For these reasons, we are constructing and testing several new Gateway-compatible vectors for expression of carboxy-terminal tags, including 1) 2X strep tag-6X His; 2) calmodulin binding protein (CBP)-3X FLAG; and 3) CBP-2X Protein A. In addition, two TEV protease sites are included in these constructs to facilitate more efficient elution from the first round of purification. To aid in detection and also potentially for use as another affinity capture moiety, these constructs also contain a tetracysteine tag that can be recognized by FlAsH™ reagents. Calmodulin binding protein, which has reversible binding, and the Strep tag, which can be competitively eluted, were chosen based on the requirement for native elutions.

To facilitate reductions in sample amounts and to increase throughput, we are exploring optimizations to our high performance liquid chromatographic (HPLC) separations and our mass spectrometric data acquisition equipment and parameters. Comparisons are underway between a more traditional three dimensional ion trap (ThermoFinnigan LCQ) and a newer linear ion trap (ThermoFinnigan LTQ). These include comparisons between data acquisition rates, sensitivity, and effective dynamic range. HPLC optimizations are being performed in parallel in order to both optimize data acquisition duty cycles and take advantage of differences in speed between the two instruments. These parallel optimizations will provide increases in throughput while maintaining or even increasing the dynamic range and sensitivity of our analysis pipeline.

In addition to affinity isolation and mass spectrometric characterization, we can identify and confirm predicted protein interactions using a live cell imaging-based assay that exploits specific localization patterns in cells. This assay involves co-expression of two fusion proteins in *Escherichia coli*. The first protein of interest is directed to the cell poles by fusion to DivIVA and the second protein of interest is fused to GFP. A direct interaction between the two proteins results in recruitment of the GFP-fusion protein to the poles. Importantly, this assay can be used to test interactions among both soluble

* Presenting author

and integral membrane proteins and is amenable to automation due to its rapidity, small scale, and ease of interpretation. To facilitate rapid analysis, we have also developed an automated image analysis algorithm to calculate the presence and location of GFP-fusion proteins in *E. coli* cells. Features such as cell number, diameter, area, and number of GFP-fusion protein localization sites are extracted from each image and used to relay quantitative values that aid in the scoring of positive interactions. This assay facilitates the directed analysis of protein interactions in live bacterial cells with the added benefit of amenability to automation.

Membrane-bound proteins are integral to many biologically active complexes, yet traditional isolation and purification of these proteins is usually tedious and inefficient. In order to make such purifications more routine, we have developed a co-fractionation strategy to localize and separate complexes under native states, followed by direct MS/MS analysis of the digested protein fractions. In this strategy we have tried to minimize dissociation between subunits, and thus loss of previously unknown subunits. For demonstration, clones of recombinant His-tagged ATP synthase were expressed in *Shewanella oneidensis* MR-1. Membrane proteins were solubilized and separated under native conditions in order of ionic strength on a Mono Q column. Fractions collected were trypsin digested and analyzed by LC MS/MS (Thermo-Finnigan LCQ). Results revealed that not only the ATP synthase subunits were eluted in common fractions, but also that proteins within other complexes were co-eluted at different ionic strengths, suggesting the presence of intact protein complexes. In parallel, we detected in-gel ATP hydrolysis approximately at the molecular size of the synthase complex (> 450 kD). Two dimensional electrophoresis images of the dissected gel show subunits ranging from 15 kD to 60 kD in size. These data demonstrate that co-fractionation and electrophoretic separation of membrane proteins coupled with mass spectrometric analysis is a valid and rapid way to analyze intracellular protein complexes.

Finally, in order to increase overall throughput, automation protocols for cloning, streaking, re-arraying, and purification steps are in place and automated affinity isolation protocols are being tested.

44 GTL

Genome-Wide Identification of Localized Protein Complexes in *Caulobacter*

P. Viollier[1]* (patrick.viollier@case.edu), J. Werner,[2] S. Pritchard,[1] E. Chen,[2] E. Huitema,[1] L. Shapiro,[3] and **Z. Gitai**[2] (zgitai@princeton.edu)

[1]Department of Molecular Biology and Microbiology, Case Western Reserve University, Cleveland Ohio; [2]Department of Molecular Biology, Princeton University, Princeton, New Jersey; and [3]Department of Developmental Biology, Stanford University, Stanford, California

Project Goals: With the availability of > 600 fully sequenced bacterial genomes, systematic approaches hold the key to exploiting our new found wealth of biological information. It is now clear that multienzyme complexes occupy discrete subcellular positions, suggesting also the existence of specialized catabolic or anabolic multi-enzyme complexes or enzymatic centers with other distinct metabolic functions. Towards the goal of assimilating a comprehensive 4D-topographic map of all co-localized proteins and protein complexes within a bacterial cell, we have embarked on a genome-wide fluorescence-based localization screen to determine the subcellular position of the finite set of proteins encoded in the genome of the Gram-negative, alpha-proteobacterium *Caulobacter crescentus*. In addition to determining the subcellular distribution of each

protein within the cell, the ability to synchronize *Caulobacter*, will allow U.S. to plot this 3D-data set into the fourth dimension (time), determining the dynamics of localization as cells progress through cell cycle. Two complementary strategies are underway by the Gitai (Princeton University) and Viollier (Case Western Reserve University) laboratories to achieve this objective.

With the availability of > 600 fully sequenced bacterial genomes, systematic approaches hold the key to exploiting our new found wealth of biological information. A major objective towards the feasibility of using bacterial cells in bioremediation and as alternate fuel sources is not only to outline protein-protein interactions on a genome-wide scale, but also determine where and when these interactions occur within the bacterial cell. It is now clear that multienzyme complexes occupy discrete subcellular positions, suggesting also the existence of specialized catabolic or anabolic multi-enzyme complexes or enzymatic centers with other distinct metabolic functions. These centers might provide the high enzymatic densities needed to facilitate maximum turn-over rates and/or they might also be important in channeling the substrates and products for ensuing reactions to the location where the reactions must take place. Towards the goal of assimilating a comprehensive 4D-topographic map of all co-localized proteins and protein complexes within a bacterial cell, we have embarked on a genome-wide fluorescence-based localization screen to determine the subcellular position of the finite set of proteins encoded in the genome of the Gram-negative, alpha-proteobacterium *Caulobacter crescentus*. In addition to determining the subcellular distribution of each protein within the cell, the ability to synchronize *Caulobacter*, will allow U.S. to plot this 3D-data set into the fourth dimension (time), determining the dynamics of localization as cells progress through cell cycle. Two complementary strategies are underway by the **Gitai** (Princeton University) and **Viollier** (Case Western Reserve University) laboratories to achieve this objective.

The Gitai laboratory is using the Gateway system from Invitrogen (www.invitrogen.com) to develop a library of fluorescently-tagged fusion proteins. There are several advantages to this approach: (1) All predicted orfs can be systematically labeled; (2) Since this is a targeted labeling scheme, we can prioritize the order of labeling and analysis of genes to investigate genes of particular interest first; (3) The number of proteins to screen for localization properties is constrained (about 3,800 for *Caulobacter*); (4) The Gateway system is modular—once entry clone are constructed, the gene of interest can be easily subcloned into a wide variety of Destination Vectors. To date, such entry clones have been generated for over 3200 genes (~85% of the predicted proteome). Roughly 2200 (~60%) of these genes have been mobilized into a destination vector that allows for their expression in *Caulobacter* as C-terminal mCherry fusions. The imaging and analysis of this first draft of the *Caulobacter* localizome is currently underway.

In a complementary approach, the Viollier laboratory has engineered random libraries of Caulobacter strains expressing protein or protein fragments fused to a fluorescent reporter (GFP or mCherry). These strains are being examined by fluorescence microscopy for fusions that are localized to specific cellular positions. There are several advantages to this approach: (1) It is unbiased; (2) It will produce localized protein fragments that will shed light on localization determinants in the complete protein; and (3) It is fast. After examining ~ 15,000 strains, approximately 0.2-3 % of the strains were found to express localized proteins, ranging from regulatory proteins to metabolic enzymes. These include several previously identified polarly-localized proteins, such as the CheR chemotaxis methylase, the DivL tyrosine kinase and the TipN polarity marker and metabolic enzymes (e.g amino acid and TCA cycle metabolism) that were also found to exhibit preferential localization to certain subcellular sites.

Follow-up studies are now underway in the Gitai and Viollier labs to elucidate how they are localized and the physiological consequences of mislocalizing them. Such studies have led to important advances on the physiological roles of these proteins and subcellular organization in *Caulobacter*

* Presenting author

(Viollier et al. 2002; Viollier et al. 2002; Gitai et al. 2004; Dye et al. 2005; Gitai et al. 2005; Huitema et al. 2006; Kim et al. 2006). In addition, the localized fusion proteins generated by both of these approaches will subsequently be systematically analyzed using an automated mislocalization screen conducted by the **Shapiro** and **McAdams** laboratories (Stanford University) in an effort to isolate *trans*-determinants for localization. These studies complement our previous experiments that charted the chromosome layout and dynamics in live *Caulobacter* cells. Together our efforts present important strides towards defining the 4-dimensional macromolecular organization of a bacterial cell using schemes that are broadly applicable to any bacterial cell of interest.

References

1. Dye, N. A., Z. Pincus, J. A. Theriot, L. Shapiro and Z. Gitai (2005). "Two independent spiral structures control cell shape in Caulobacter." Proc Natl Acad Sci U S A 102(51): 18608-13.
2. Gitai, Z., N. Dye and L. Shapiro (2004). "An actin-like gene can determine cell polarity in bacteria." Proc Natl Acad Sci U S A 101(23): 8643-8.
3. Gitai, Z., N. A. Dye, A. Reisenauer, M. Wachi and L. Shapiro (2005). "MreB actin-mediated segregation of a specific region of a bacterial chromosome." Cell 120(3): 329-41.
4. Huitema, E., S. Pritchard, D. Matteson, S. K. Radhakrishnan and P. H. Viollier (2006). "Bacterial birth scar proteins mark future flagellum assembly site." Cell 124(5): 1025-37.
5. Kim, S. Y., Z. Gitai, A. Kinkhabwala, L. Shapiro and W. E. Moerner (2006). "Single molecules of the bacterial actin MreB undergo directed treadmilling motion in Caulobacter crescentus." Proc Natl Acad Sci U S A 103(29): 10929-34.
6. Viollier, P. H., N. Sternheim and L. Shapiro (2002). "A dynamically localized histidine kinase controls the asymmetric distribution of polar pili proteins." EMBO J 21(17): 4420-8.
7. Viollier, P. H., N. Sternheim and L. Shapiro (2002). "Identification of a localization factor for the polar positioning of bacterial structural and regulatory proteins." Proc Natl Acad Sci U S A 99(21): 13831-6.

45 GTL

The Structure and Function of the *Caulobacter* MreB Actin-Like Cytoskeleton

N. Dye,[1,2]* M. Mielke,[3] Z. Pincus,[2] J. Theriot,[2] L. Shapiro,[1] and **Z. Gitai**[3] (zgitai@princeton.edu)

[1]Department of Developmental Biology, Stanford University, Stanford, California; [2]Department of Biochemistry, Stanford University, Stanford, California; and [3]Department of Molecular Biology, Princeton University, Princeton, New Jersey

Project Goals: The bacterial MreB cytoskeleton is an actin-like structure that represents an essential integrator of spatial and temporal cellular information. Despite its importance, however, the mechanisms by which MreB regulates processes such as cell morphogenesis, chromosome segregation, and protein localization remain unknown. Consequently, our objective is to determine the structure, function, and regulation of the MreB cytoskeleton. These studies will both illuminate broadly conserved processes that are essential to the survival of all bacterial species, and serve as a road-map for the mechanistic dissection of additional dynamic protein structures in bacteria.

The bacterial MreB cytoskeleton is an actin-like structure that represents an essential integrator of spatial and temporal cellular information. Despite its importance, however, the mechanisms by which MreB regulates processes such as cell morphogenesis, chromosome segregation, and protein localization remain unknown. Consequently, our objective is to determine the structure, function, and

regulation of the MreB cytoskeleton. These studies will both illuminate broadly conserved processes that are essential to the survival of all bacterial species, and serve as a road-map for the mechanistic dissection of additional dynamic protein structures in bacteria.

To probe the structure and organization of MreB, we are developing novel methods of analyzing its assembly and dynamics, both *in vivo* and *in vitro*. By imaging and tracking single molecules of fluorescently-labeled MreB, we have been able to visualize MreB kinetics in living cells (Kim et al. 2006). From these data we determined that MreB monomers treadmill through individual filaments in a polar manner, but that the polarity of each individual filament appears independent of the overall cellular polarity. We have also used these tracking data to model MreB's average filament length and polymerization and depolymerization rates Kim, 2006 #276}. These results suggest that the previously-observed MreB helical ultrastructures consist of multiple short, bundled filaments. To directly validate these models, we are now collaborating with the Ellisman (UCSD) and Larabell (LBNL) groups to develop EM- and XM-compatible strategies for ultra-high-resolution-imaging of MreB and MreC structures.

To identify proteins that function with MreB, we first took a candidate approach, examining proteins that are co-conserved with MreB. By combining mutant and localization analysis, we determined that the cytoplasmic MreB cytoskeleton functions in parallel to a periplasmic helical structure made up of the MreC protein (Dye et al. 2005). By developing a novel computational method for analyzing cell shapes, we also determined that MreB and MreC collaborate to determine proper morphology by regulating the proper localization of a transmembrane peptidoglycan rearrangement enzyme, Pbp2 (Dye, Pincus et al. 2005). Our shape analysis software should be widely applicable to other researchers studying irregular cell morphologies.

Currently, we are using several genetic and biochemical approaches to identify and characterize additional upstream regulators and downstream mediators of MreB function. We have succeeded in assembling filaments of purified MreB *in vitro*, and we are now characterizing the mechanistic details of MreB polymerization. Such *in vitro* polymerization of fluorescently tagged MreB is providing a method to assay candidate MreB regulators. Together, these studies will help U.S. understand a fundamental and conserved macromolecular complex, while serving as a platform for the development of new methods that will be broadly applicable to other bacterial systems of interest to the DOE.

References

1. Dye, N. A., Z. Pincus, J. A. Theriot, L. Shapiro and Z. Gitai (2005). "Two independent spiral structures control cell shape in *Caulobacter*." *Proc Natl Acad Sci U S A* **102**(51): 18608-13.
2. Kim, S. Y., Z. Gitai, A. Kinkhabwala, L. Shapiro and W. E. Moerner (2006). "Single molecules of the bacterial actin MreB undergo directed treadmilling motion in *Caulobacter crescentus*." *Proc Natl Acad Sci U S A* **103**(29): 10929-34.

 * Presenting author

46 GTL

EM Tomography Enhancements

Fernando Amat[1]* (famat@stanford.edu), Farshid Moussavi[1]* (farshid1@stanford.edu), Kenneth H. Downing,[3], Mark Ellisman,[4] Luis R. Comolli,[3] Albert Lawrence,[4] Mark Horowitz,[1]* and **Harley McAdams[2]** (hmcadams@stanford.edu)

[1]Electrical Engineering Department and [2]Developmental Biology Department, Stanford University, Stanford, California; [3]Lawrence Berkeley National Laboratory, Berkeley, California; and [4]National Center for Microscopy and Imaging Research, University of California, San Diego, California

Project Goals: Development of high-throughput methods to identify and characterize spatially localized multiprotein complexes in bacterial cells.

Electron microscope tomography is a powerful tool for studying the 3D structure and organization of biological specimens. However, it presents particular challenges for attaining high enough resolution. One such limitation is the inability to collect high resolution images of sections at high tilt due to apparent specimen thickening that occurs as they are tilted. This effect produces blurred images due to chromatic aberrations at high tilt, in addition to the so-called "missing wedge" of data, which together contributes to a significant reduction in resolution of tomographic reconstructions, especially in the z-axis. We have been pursuing several approaches in order to reduce these effects, including development of most-probable loss (MPL) energy filtering electron tomography [1], but are still ultimately constrained by the specimen geometry. To augment this MPL strategy we have developed a method allowing the fabrication of "prismatic" sections of resin embedded microbial specimens for ultra high tilt electron tomography. The procedure for preparing these and the associated tomographic volumes will be described.

Another obstacle to high resolution is alignment accuracy. This is especially true in the case of cryo EM tomography, which enables the study of cells in close to their "native" environment. Each image in a cryo EM tilt series is quite noisy, since the total electron dose through the tomographic tilt series must be constrained to limit structural damage to the cell. Even with gold markers added to the sample, robust automatic alignment of the cryo EM image data for reconstruction remains difficult, and manual intervention is required. We address this problem by leveraging recent work in probabilistic analysis, and have constructed a prototype alignment system using *Markov random fields (MRF's)* and robust optimization for alignment of tilt-series. This fully automatic alignment will become more critical as faster and better systems for automatic data acquisition are being developed. Our goal is to eliminate the alignment bottleneck in tomographic imaging. We are also integrating these methods with recent approaches that allow for correction of projection errors due to specimen changes during a tilt series and distortions due to properties of the electron optics [2].

With markers, there are three basic steps required to align the cryo EM dataset: marker feature identification, correspondence and tracking of these features throughout the image set, and projective model estimation from these feature tracks. Typically, automatic tracking makes many errors, which in turn cause the projection model estimation to fail. In our framework, we focus on reduction of tracking errors through the use of contextual information, as well as making the projective model estimation robust to any remaining tracking errors. Feature correspondence and tracking are accomplished at local and global levels. Local correspondence is between two images, and is accomplished by treating the set of markers as a Markov random field (MRF), that is, a set of variables that depend on each other only through their neighbors. We use mutual information and the relative geometric positions of pairs of markers to set the initial estimates for local correspondence, and use a standard

loopy belief propagation algorithm to estimate the most likely correspondences between image pairs. Global correspondence is achieved by combining the results of local correspondence in a tree-based comparison scheme with redundancy to form robust track estimates.

Errors in the tracks are possible due to feature location mistakes as well as inaccurate inference results. Therefore, the projective model estimation uses a robust fitting method as opposed to least squares that is tolerant to outliers. After we have an estimate of the projective model, the model is iterated using expectation maximization (EM) to re-estimate perceived outliers with improved reprojection data from the current model. This iteration is performed as many times as necessary before a stopping criterion is satisfied. In sample cases, we find that only a small number of iterations is needed (often only one).

This robust alignment framework has allowed U.S. to fully automatically recover dozens of contours (both complete and piecewise) with subpixel accuracy from several challenging cryo datasets of the bacterium *Caulobacter crescentus*. The results were used to create 3D reconstructions comparable to results previously obtainable only by extensive manual intervention.

All the techniques described above are implemented in the software package Robust Alignment and Projection Estimation for Tomographic Reconstruction (RAPTOR). This software will be made generally available.

References

1. Bouwer et al. (2004) J Struct Biol. 148(3):297-306.
2. Lawrence et al. (2006) J Struct Biol. 154(2):144-67.

47 GTL

Automated Screen for Identification of Mislocalization and Morphological Mutants

G. Bowman, N. Hillson, M. Fero, S. Hong, L. Shapiro, and **H. H. McAdams*** (hmcadams@stanford.edu)

Department of Developmental Biology, Stanford University, Stanford, California

Project Goals: Development of high-throughput methods to identify and characterize spatially localized multiprotein complexes in bacterial cells.

Many dynamically localized bacterial proteins have been identified over the last few years by fluorescence microscopy, and in many of these cases the localization is essential for proper cell function. The number of localized proteins will vary across organisms. In *Caulobacter crescentus*, roughly 10% of expressed proteins are localized at some time in the cell cycle. Although the functions of many localized proteins have been extensively studied, there is limited understanding of the pathways and dependencies that produce localization or of the protein domains required for localization. Identification of these pathways and mechanisms involves laborious characterization of mutant strains. This is traditionally a "low throughput" process, particularly where the mutant phenotypes can only be detected by microscopic examination. We have developed a robotic screen to massively accelerate identification of mislocation mutants using automated microscopy and computational image analysis.

* Presenting author

The localization mutants identified are then characterized and analyzed by conventional genetic and biochemical methods.

Our automated, high-throughput screen process can identify mutant strains that mislocalize both essential and non-essential proteins that normally have a distinctive spatial or temporal localization pattern. Development of this screening technology is a collaboration between the McAdams and Shapiro labs at Stanford, the Viollier lab at Case Western, and the Gitai laboratory at Princeton.

There are two stages to the screen process. First, we identify the proteins that exhibit a temporal or spatial localization pattern and characterize that localization pattern. This is being done in collaboration with the Viollier laboratory at Case Western University and the Gitai laboratory at Princeton University. Second, for a subset of the localized proteins, we use genetic screening to identify mutations that cause localization defects (in timing or placement). For each target localized protein, we generate strains with location defects. We use transposon mutagenesis for the nonessential genes. The most challenging mutants, of course, are those targeting proteins whose proper location is essential. We generate temperature-sensitive mutants for these strains. Location defects may involve complete failure to localize or localization at the wrong time or place. The screen will identify both intragenic mutations that indicate portions of the protein required for its cellular localization, and extragenic mutations in genes encoding interacting proteins required for either specific or general localization processes.

We use the automated microscopy system to examine thousands of the mutagenized strains to identify those with location defects. We first spot the mutagenized strains on a novel microscope slide design to obtain an array of samples on small agar pads under a cover slip. Fluorescent micrographs of the resulting cell sample arrays are taken rapidly with the computer controlled microscope. The rectangular arrays of agar pads provide a specific row-column address on the slide that is correlated with the source wells in the 96-well storage. We have developed automatic image processing algorithms to find, orient, measure, and classify cells in the microscopic images. This brute force, yet totally automated, system for finding localization mutants is generally applicable to any protein whose GFP-tagged variant shows a distinctive localization pattern in the cell.

An integral part of our screen methodology involves detailed computational analysis of the size and shape of collections of cells in the microscope image field. The program we developed for the image analysis automatically extracts individual cell images from a field of cell images. The properties of the cell shapes and localized protein positions are then used to classify the cells according to stage of cell development and localization pattern. The characteristics of the sample cells in the mutagenized sample are then compared to characteristics of wild type cell populations. Sorting by cell cycle stage is necessary because many proteins are transiently localized at a distinct time in the cell cycle. Although our primary focus is on identification of mislocation mutants, the image classification analysis also identifies mutant strains with defects in cell shape or in cell cycle progression. This enables identification of cytoskeleton or cell decision defects as a valuable byproduct result.

The key requirement for the screen is the ability to examine very large numbers of mutagenized strains by fluorescent microscopy. Key characteristics of the fluorescent microscope and associated acquisition software are automated X-Y-Z stage control, automated switching from transmission microscopy modes such as phase contrast or DIC to fluorescence imagery, automated time lapse imaging, and fully automatic focusing. In addition, the temperature of the sample and stage are monitored and controlled via an environmental enclosure. Images are automatically acquired using a megapixel cooled CCD camera and transferred to an image repository where they are queued for image analysis. Automation of computer control enables thousands of strains to be screened rapidly with minimal manual involvement in the microscopy. We can use a loose tolerance for false positives

in the automated image analysis, since the cell images that the computer analysis categorizes as probable mutants are then examined by eye.

For our initial mutant search, we constructed a strain with three localized proteins labeled with distinct fluorescent tags: PleC, which localizes to the swarmer pole, was labeled with YFP; DivJ, which localizes to the stalked pole, was labeled with mCherry; and ZapA, which localizes to the division plane, was labeled with CFP. The use of the multi-labeled strain enabled a simultaneous search for strains with mislocation of any of the three labeled proteins, so that the mutant yield is higher.

48 GTL

Methods for *in vitro* and *in vivo* Imaging of Protein Complexes

Huilin Li* (hli@bnl.gov), James Hainfeld* (hainfeld@bnl.gov), Minghui Hu, Michael Mylenski, Kevin Ryan, Luping Qian, Raymond P. Briñas, Elena S. Lymar, and Larissa Kusnetsova

Biology Department, Brookhaven National Laboratory, Upton, New York

Project Goals: To develop new high resolution imaging method and to apply the method for imaging protein complexes in biological systems that is relevant to DOE mission.

The structures of biological molecular assemblies and their locations inside the cell are keys to understanding the function. Our *in situ* bi-mode cryo-electron tomography and site-specific labeling method, which takes advantage of ultra-structural visualization capability of the cryo-TEM and the heavy metal cluster label detection capability of the cryo-STEM, holds the potential for simultaneous three-dimensional structural visualization and protein mapping.

Developing the cryo-STEM tomography capability on a commercial Jeol 2010F cryo-TEM/STEM microscope.

Our approach requires low dose imaging and tomographic capability in both TEM and STEM modes. A modern commercial transmission electron microscope comes with these capabilities only in TEM mode, but not in STEM mode. In last year, we developed a Gatan Digital Micrograph Plug-In that we called it STEMan that enable low dose image acquisition. During this year, we added STEM tomography capability, and the program is now called AuotSTEM for automated scanning transmission electron microscopy tomography (Fig. 1). AutoSTEM was based on scripts created using Digital Micrograph (DM), since we use the Gatan STEM acquisition device DigiScan. Jeol's FasTEM Communication Kit (FTCOMM) and Digital Micrograph's Software Developers Kit (DMSDK) were essential in creating new functions to implement auto-tracking and auto-focusing. Overall, we developed five DM scripts and one Microsoft's Visual Studio–based dynamic link library (DLL). The DLL communicates between the microscope and DM. The Auto-focusing part utilized an image-gradient based algorithm. The computation for auto-tracking and auto-focusing is primarily based on the DM scripts. The program is completed but further testing is needed.

* Presenting author

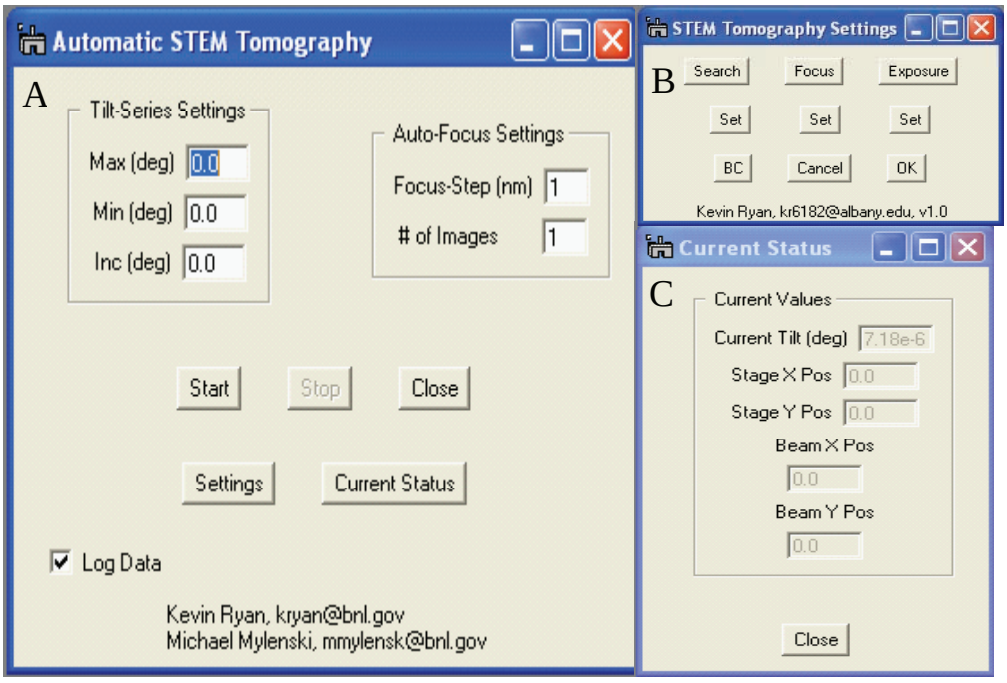


Fig. 1. GUIs for AuotSTEM. (A). The main Graphical User Interface (GUI) in Digital Micrograph for automated STEM tomography. It controls the auto-tilting, auto-focusing parameters, the starting and stopping of the tilt series, and access to the settings and current status GUIs. (B). GUI for adjusting microscope settings for auto-tracking and auto-focusing. (C) GUI when invoked displays the current microscope settings, such as specimen tilt angle, the current stage and current beam positions.

Nanoparticle assembly of bacterial proteins and complexes and electron microscopy – Potential use for alternative energy production.

We have used site-specific binding gold nanoparticles to assemble proteins and complexes into higher order structures that could be useful for alternative energy production and visualized them by EM. For example, redox enzymes can oxidize ethanol to produce free electrons which, if harnessed, would directly produce electricity. In order to do this, the enzymes must be assembled on an electrode and the flow of electrons efficiently coupled to it. This was shown to be possible for glucose oxidase using gold nanoparticles: the electron would normally be received by O_2 in solution, but could be routed to a metal electrode by attaching the enzyme and providing a conductive path via the nanoparticle to the metal surface (Xiao Y, Patolsky F, Katz E, Hainfeld JF, Willner I: Plugging into Enzymes: nanowiring of redox enzymes by a gold nanoparticle. Science, 299, 1877, 2003.).

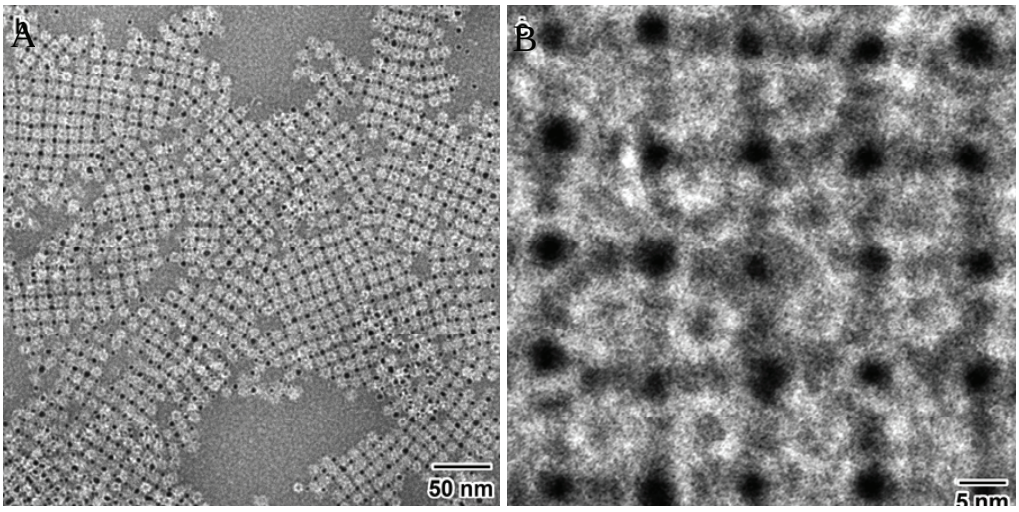


Fig. 2. The bacterial 20S proteasomes organized into 2-D arrays using 4.4 nm gold nanoparticles (black dots). (A) low mag and (B) high mag.

One obstacle to producing efficient enzyme electrodes is to coat the electrode with the redox enzyme in a functional monolayer. Here we have taken bacterial enzyme complexes and shown by electron microscopy that large monolayers could be assembled by use of gold nanoparticles functionalized to "glue" the enzymes together into an oriented and ordered sheet. This sheet was shown to be strong enough to transfer onto a carbon film, the EM grid (Fig.2). The protein used was *Mycobacterium tuberculosis* 20S proteasomes which were expressed with a 6x-His tag. They were organized into arrays by binding to multifunctional 4.4 nm gold nanoparticles derivatized with the nickel nitrolotriacetic acid group (Ni-NTA). **Using this newly developed nanotechnology, it should be possible to create ordered redox enzyme electrodes leading to more efficient energy conversion to electricity of biofuels.**

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

49 GTL

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

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

[1]Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, Georgia; [2]Department of Biology, California Institute of Technology, Pasadena, California; and [3]Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania

Project Goals: The goal of this DOE/GTL project is to develop quantum-dot (QD) based strategies for imaging and identification of individual proteins and protein complexes in microbial cells with high specificity and sensitivity. The multifunctional quantum-dot bioconjugates consist of quantum dots of 2-6 nm in size encapsulated in a phospholipid micelle, and delivery peptides and protein targeting ligands conjugated to the surface of the QDs. Once the QD bioconjugates are

* Presenting author

internalized into microbial cells by the peptide, the adaptor molecules on the QD surface bind to specific target proteins or protein complexes that are genetically tagged. Optical imaging is used to visualize the localization, trafficking and interaction of the proteins, resulting in a dynamic picture but with a limited spatial resolution (~200 nm). The same cells are then imaged by EM to determine their detailed structures and localize the target proteins to ~4 nm resolution. For each protein or protein complex, selected tags will be tested to optimize the specificity and signal-to-noise ratios. This innovative molecular imaging approach integrates peptide-based cellular delivery, protein targeting/tagging, light microscopy and electron microscopy, and will have a wide range of biological applications relevant to the GTL program at DOE.

We have been developing quantum-dot (QD) based strategies for imaging and identification of individual proteins and protein complexes in microbial cells. Currently, there is a lack of novel labeling reagents for visualizing and tracking the assembly and disassembly of multi-protein molecular machines. There is no existing method to study simultaneous co-localization and dynamics of different intra-cellular processes with high spatial resolution. As shown in Figure 1, the multifunctional quantum-dot bioconjugates we develop consisting of a quantum dot of 2-6 nm in size encapsulated in a phospholipid micelle, with delivery peptides and protein targeting ligands (adaptors) conjugated to the surface of the QD through a biocompatible polymer. After internalization into microbial cells, the adaptor molecules on the surface of QD bioconjugates bind to specific target proteins or protein complexes that are genetically tagged. Optical imaging is used to visualize the localization, trafficking and interaction of the proteins, resulting in a dynamic picture but with a limited spatial resolution (~200 nm). The same cell is imaged by EM to determine their detailed structures and localize the target proteins to ~4 nm resolution. For each protein or protein complex, selected tags are tested to optimize the specificity and signal-to-noise ratios of protein detection and localization. This innovative molecular imaging approach integrates peptide-based cellular delivery, protein targeting/tagging light microscopy and electron microscopy.

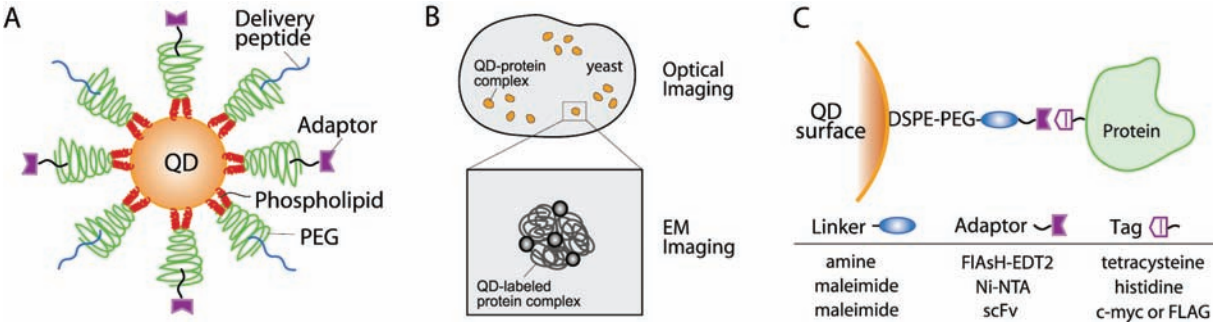


Figure 1. (A) Schematic illustration of a multifunctional quantum dot bioconjugate consisting of encapsulated QD with targeting adaptor and delivery peptide on its surface; (B) correlated optical and EM imaging of the same cell gives both temporal and spatial information on a protein complex; (C) possible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

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

To target semi-conductor quantum dots (QDs) to specific intracellular proteins, we constructed fusion proteins including the commercially available SNAP tag. The SNAP tag is a 22 kDa protein that is an engineered form of the human O6-alkylguanine-DNA-alkyltransferase (AGT). The SNAP protein is also able to transfer residues from para-substituted benzylguanines, resulting in the covalent attachment of the substituted group (such as a QD) to the SNAP tag. As a model system, SNAP-DsRed-Monomer-Actin fusion proteins were expressed in *E.coli* and purified in order to perform *in vitro* polymerization experiments to prove the functionality of the fusion proteins, which was assessed both by polymerizing actin and then adding QDs for tagging of the filaments and by attempting to polymerize the actin already labeled with the QDs. The use of the DsRed-Actin fusion allowed for continuous monitoring of both the actin and the QDs without the need for fixing and staining. In addition, the pore-forming bactotoxin streptolysin-O (SLO) was used to deliver benzylguanine-conjugated QDs into 3T3 cells transfected with pSNAP-DsRed-Monomer-Actin. The feasibility of using this covalent method to label intracellular proteins in live cells was assessed.

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

We are advancing electron tomography as a promising new tool to image protein complexes both *in-vitro* and *in-vivo* within small microbial cells. A new helium-cooled, 300kV, FEG, "G2 Polara" FEI TEM at Caltech was used to image purified protein complexes, viruses, and whole bacterial cells. We pioneered the use of a new "flip-flop" cryorotation stage that allows dual-axis cryotomography, and developed a simple Perl-based system for distributed computation to handle the massive image processing demands that arise from imaging intact bacteria in 3D. These technological advances have allowed U.S. to visualize directly cytoskeletal elements within small microbial cells and the domain structure of purified multienzyme complexes, both are key imaging goals of the Genomics:GTL program. For example, using electron cryotomography of whole cells, we revealed the *in situ* structure of the complete flagellar motor from the spirochaete *Treponema primitia* at 7 nm resolution. Twenty individual motor particles were computationally extracted from the reconstructions, aligned and then averaged. The stator assembly, revealed for the first time, possessed 16-fold symmetry and was connected directly to the rotor, C ring and a novel P-ring-like structure. The unusually large size of the motor suggested mechanisms for increasing torque and supported models wherein critical interactions occur atop the C ring, where our data suggest that both the carboxy-terminal and middle domains of FliG are found.

We have successfully integrated quantum dots into *Dictyostelium* through culturing them simultaneously with bacteria, on which they are feeding. After incubation and both fluorescent and confocal microscopy imaging, the *Dictyostelium* reveal a distribution of quantum dots indicating that they are dispersed throughout the cell cytoplasm. This is compared with the control cells that have not been incubated with quantum dots and further with cell studies where endosomes reveal aggregated fluorescent patterns at earlier time periods. We have shown specific live cell labeling of *Dictyostelium* tagged with quantum dots that are targeted for F-actin using phalloidin. The fluorescent patterns are similar to the patterns for *Dictyostelium* that is immunofluorescently labeled with FITC-phalloidin.

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

* Presenting author

50 GTL

Correlated Light and Electron Microscopy of Protein Complexes in *Caulobacter crescentus*

Guido M. Gaietta[1]* (ggaietta@ncmir.ucsd.edu), Thomas J. Deerinck,[1] Grant Bowman,[2] Yi Chun Yeh,[2] Luis R. Comolli,[3] Lucy Shapiro,[2] Harley McAdams,[2] and **Mark H. Ellisman**[1]

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

Project Goals: Adaptation of a molecular tagging approach to correlated light and electron microscopy for 3D analysis of protein complexes in *Caulobacter crescentus*.

Many regulatory proteins and protein/DNA complexes in *Caulobacter crescentus* are found in specific locations with noted variation related to the stage in the cell cycle. A main objective of the *Dynamic spatial organization of multi-protein complexes controlling microbial polar organization, chromosome replication, and cytokinesis* GTL project is to adapt methods for correlated LM and EM imaging to the analysis of these structures in *Caulobacter*. Our efforts have been focusing on the use of the tetracysteine/biarsenical molecular tagging system, developed in collaboration with Professor Roger Tsien's group at UCSD. Recombinant probes for proteins in key subsystems of *Caulobacter* are being generated and used as test bed to develop this approach, which allows U.S. to achieve optimal preservation of the ultrastructure, sensitive localization of the target complexes and their visualization in 3D reconstructions. The following table presents some of the strains currently under analysis:

Protein	Tag	Localization
FtsZ	GFP-4C, YFP-4C	Division plane
FtsK	GFP-4C	Division Plane
HU	GFP-4C	Chromatin
MipZ	GFP-4C, YFP-4C	Cell pole
MreC	4C	Cytoskeleton
MreB	4C	Cytoskeleton
LacI	GFP-4C, CFP-4C	Chromatin
McpA	GFP-4C	Cell Pole
CpaF	4C	Cell Pole
GmpA (Snc04)	GFP-4C	Cell Pole

As displayed in the table, recombinant forms of the target proteins often carry combinations of intrinsically fluorescent proteins (FP; GFP and its derivatives) tagged at the carboxyl terminus with the small tetracysteine motif FLNCCPGCCMEP (4C). The resulting protein is visible at the light level by direct excitation of FP and, upon labeling of the tetracysteine motif, of the biarsenical compounds FlAsH and ReAsH. ReAsH is also used to trigger the photoconversion of diaminobenzidine (DAB) by either direct excitation or by FRET from the neighboring FP, hence generating a precipitate visible at the electron microscope (upon treatment with osmium tetroxide).

We are currently developing the labeling protocol using the pXyl-GmpA strain, in which we can achieve robust expression of tetracysteine tagged GmpA protein using a multi-copy cassette under control of the xylose-dependant promoter pXyl. GmpA is a newly identified protein that localizes to the cell poles throughout the cell cycle of *Caulobacter*, and is essential for normal cell division.

When we induce high levels of protein expression in the pXyl-GmpA strain, the GmpA protein forms a large plug at the cell pole that displaces the normal cytoplasm. This enlarged region of GmpA accumulation has the effect of increasing the area of the cell pole, as the localization patterns of other polar proteins are extended to co-localize with the GmpA plug.

Using these enlarged polar features, we have shown that it is possible to apply the ReAsH labeling method to observe protein localization at high resolution at the cell poles. To do this, the labeling process was adapted from that used in eukaryotic systems, and streamlined so that *Caulobacter* cells could be easily stimulated, stained and prepped for photoconversion over a ~4 hour-long period. After stringent washes with a competing dithiol agent, ReAsH-labeled bacteria were pelleted, re-suspended and plated onto polyethylenimine-coated Mat-Tek dishes and, for optimal preservation of the bacterium ultrastructure, fixed in a glutaraldehyde/acrolein mixture. The photoconversion reaction is driven by exciting ReAsH by FRET from GFP, hence increasing the overall absolute contrast of the biarsenical, bringing it closer to that of GFP. Furthermore, FRET excitation excluded non-specific biarsenicals because of its tight space limitation between donor and acceptor (they must be no more than 0.8 nm apart), improved the actual specificity of photoconversion and provided a highly concentrated and reproducible deposition of DAB polymers at the cell pole, where GmpA locates. Very recent advancements in our technique have increased labeling efficiency, and we are making progress towards the goal of observing GmpA localization at normal expression levels. Eventually, we will compare the localization patterns of several polar proteins to determine the three dimensional structural organization of this important cellular domain. Applying this technique to proteins that have other kinds of localization patterns (table 1) will eventually allow U.S. to observe chromosome organization, division plane assembly, cytoskeletal proteins, and more.

In order to obtain an accurate, high-resolution map of the organization of proteins in *Caulobacter*, it is imperative to optimize the preservation of cellular ultrastructure. We find that we can greatly enhance preservation by combining the aldehyde/acroelin cross-linking properties and High Pressure Freezing (HPF) and freeze substitution. This hybrid method allows the introduction of a photoconversion step to deposit DAB for subsequent osmification during the low-temperature freeze-substitution process. Electron Tomography proceeds with the resin embedded samples with preservation comparable to that obtained by HPF from live specimens in absence of chemical cross-linking.

51 GTL

Computational Analysis of the Protein Interaction Networks of Three Archaeal Microbes

Chris Ding* (chqding@lbl.gov), Chunlin Wang, and **Stephen R. Holbrook**

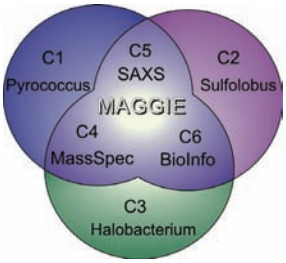
Lawrence Berkeley National Laboratory, Berkeley, California

Project Goals: MAGGIE will address immediate GTL goals by accomplishing the following three overall goals: 1) provide a comprehensive, hierarchical map of prototypical microbial protein interactions, 2) develop and apply SAXS technologies for high throughput characterizations of protein conformations, shapes and assemblies controlling microbial cell biology, and 3) create and test powerful computational descriptions for the interpretation and eventual control of protein interactions and activities controlling microbial processes. We propose to develop

* Presenting author

and implement a framework for integration, representation and analysis of proteomic data from both experimental and bioinformatics sources. This multi-scale representation proceeds from protein motifs to domains, proteins, complexes, super-complexes and pathways.

The computational identification and mapping of protein complexes and networks using genomic and experimental data is critical to our understanding of complex biological systems. In relation to the MAGGIE project to characterize the molecular machines and interactions of archaeal genomes, we have developed a software suite for analysis of their protein interaction network. These algorithms identify missing links in the network by transitive closure, extracted protein complexes as cliques, placed them into context by a hierarchical decomposition process using a distance metric and allowed for proteins to be shared between complexes. Finally, we have developed an algorithm for extracting bicliques from a data matrix and applied it to finding bicliques between protein domains and protein complexes.



Because of the abundance of experimental data, we have used yeast in development and testing of these programs, but as experiments proceed on microbial proteomes, we have begun to apply these approaches to *Pyrococcus furiosus*, *Sulfolobus solfataricus*, and *Halobacterium* NRC-1. An advantage to studying these three diverse archaea is the capability to characterize conserved components and organism specific systems of the microbial domain. This software and results drawn from analysis of these organisms will be made available to the GTL community.

52 GTL

The Use of Small Angle X-ray Scattering to Extract Low Resolution Structures and Monitor Sample Quality from Archeal Proteomes

Greg Hura[1]* (glhura@lbl.gov), Michal Hammel,[1] Susan Tsutakawa,[1] Cesar Luna-Chavez,[1] Robert Rambo,[1] Ferris Poole,[3] Francis Jenney,[3] Angeli Lal Menon,[3] Mike Adams,[3] and **John Tainer**[1,2]

[1]Lawrence Berkeley Natiobal Laboratory, Berkeley, California; [2]Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, California; and [3]Department of Biochemistry & Molecular Biology, University of Georgia, Athens, Georgia

Project Goals: Characterization of Archeal Proteomes

Small Angle X-ray Scattering (SAXS) is a high through-put technique for determining low resolution structures of macromolecular complexes. Ideal samples for SAXS data collection are 15-2μg/μL of monodispersed and homogeneous macromolecules in solution. Samples which do not fit these criterion are identifiable from the SAXS signal and therefore SAXS may be used as a quality check on various preparations of a particular macromolecule. Samples which do fit this criterion produce scattering curves which may be used to define dimensions above 10Å.

The SIBYLS beamline at the Advanced Light Source in Berkeley has been developed to conduct high through-put SAXS. Using a pipetting robot 800 samples may be collected in 40 hours. We intend to characterize native and recombinantly purified samples from organisms of GTL interest. By screening various buffer conditions we will determine stabilizing conditions for particular complexes. By screening libraries of compounds we will determine which metabolites cause con-

formational changes in various macromolecules. We will also develop and maintain a web based data base where low resolution structures of complexes are accessible to the scientific community.

A first pass has been conducted on 17 Pyrococus Furiosis (Pf) recombinant proteins which were previously prepared as part of a crystallographic structural genomics initiatives. Three such samples were successfully crystallized in that effort while 14 remained structurally uncharacterized. Of the 17 samples collected 13 were immediately amenable for shape reconstruction, the other 4 were aggregating in solution. Various preparative conditions were attempted with one of the 4 aggregates which identified a successful non-aggregating preparation.

The low resolution shapes for the 3 samples which had a crystallographically determined structure showed excellent agreement with their structures. Within the remaining 10, for which low resolution structures were determined, large multimers were identified. Many of the structures have large flexible tails which likely complicate crystallographic studies. Proteolysis fragments may be more amenable. Some of the complexes have no homology to proteins with known function. Further biochemical characterization of the most interesting shapes will continue.

53 GTL

Imaging *Caulobacter crescentus* using Soft X-Ray Tomography: A New Imaging Tool for Genomics:GTL and Bioenergy Research

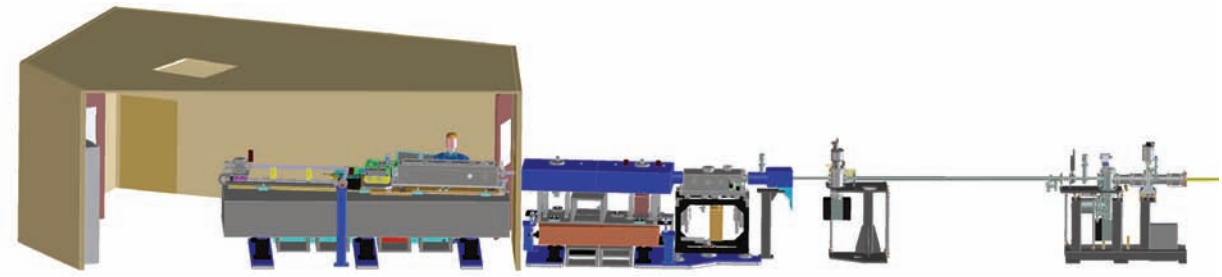
B.M. Maguire,[1,3] C.A. Tonnessen,[1,3] G. McDermott,[1,3] A.J. McDonnell,[1,3] M.A. Le Gros,[2,3] and **C.A. Larabell**[1,2,3*] (CALarabell@lbl.gov)

[1]Department of Anatomy, University of California, San Francisco, California and [2]Physical Biosciences Division and [3]National Center for X-ray Tomography, Lawrence Berkeley National Laboratory, Berkeley, California

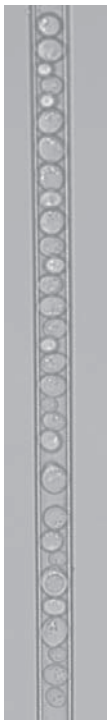
Project Goals: Charting cellular sub-structures and localizing proteins and multi-protein complexes in whole hydrated cells using X-ray tomography.

Developing a predictive systems-level understanding of an organism requires integration of data from a number of large scale '-omics' programs, such as proteomics, metabalomics, and structural genomics. An unambiguous interpretation of this data also requires detailed knowledge of the cellular locations where the molecular interactions occur during the cell cycle, or when the organism is responding to specific environmental conditions. Soft x-ray tomography is a new tool for imaging cells and for localizing labeled proteins. This new imaging technique has an inherent advantage over established imaging techniques, such as electron microscopy, in that it only requires the use of simple sample preparation protocols and results in images with a spatial resolution significantly higher than can be obtained with light microscopy. The technique also has a further advantage in that soft x-rays images revel the internal architecture of a whole cell without the need to dehydrate the cell, use fixatives, or potentially damaging contrast agents. This is, therefore, an ideal method for simultaneously visualizing phenotypic plasticity and the cellular location of critical macromolecules.

* Presenting author



CAD drawing of XM2, a new soft x-ray microscope for biological and bio-energy research at the Advanced Light Source, Berkeley



Yeast cells mounted in a thin-walled capillary tube, ready for imaging using soft x-ray microscopy

An important factor in the applicability of this new technique to Genomics:GTL and Bioenergy research is the level of sample throughput that can be sustained. Typically, a complete tomographic data set can be collected from a bacterial cell in less than three minutes. In addition, upwards of twenty cells can be stacked horizontally in a glass capillary sample mounting device. A small translation of this capillary at the end of data collection quickly brings fresh cells in to the field of view, ready for imaging. In this way, information on the sub-cellular architecture of the bacterium, and the locations where critical molecular interactions take place, can be obtained with statistical significance in a short space of time.

We are currently applying soft x-ray microscopy to the study of cell development, and formation of the polar regions in the bacteria *Caulobacter crescentus*. Towards this end, we are developing technologies for labeling proteins in this bacteria for imaging by both light and x-ray microscopy. Details of these methods will be presented on the poster

Funded jointly by the NIH and the DOE, the new soft x-ray microscope was constructed at the worlds brightest soft x-ray source, the Advanced Light Source of Lawrence Berkeley National Laboratory. The new instrument was completed at the end of 2006, and is now being commissioned. The facilities are open to use by any qualified scientist. The external user program is expected to commence, in limited numbers, in the summer of 2007.

References

1. X-ray tomography of whole cells, MA Le Gros, G. McDermott, & CA Larabell. (2005) *Current Opinion in Structural Biology*, **15**, 593-600

 * Presenting author