See addendum for additional abstracts.

Systems Biology Research Strategy and Technology Development

Genomic and Proteomic Strategies

99

GTL

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

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

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

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

Introduction

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

Approach

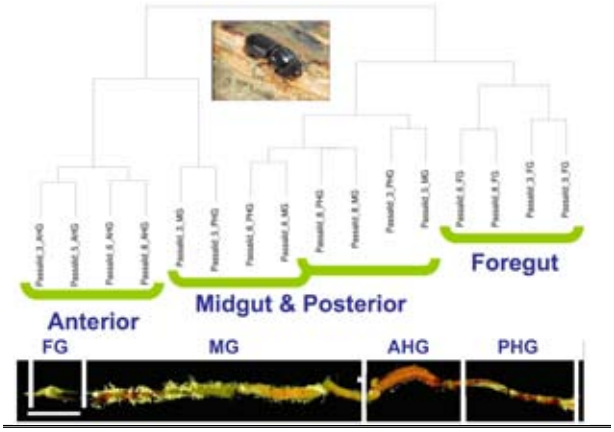
This concept involves labeling of microbial nucleic acids following incubation with a stable isotope-labeled compound (e.g ^{13}C-cellulose or $^{15}N_2$). Extracted RNA is hybridized to a newly engineered high-density oligonucleotide microarray with a conductive surface and higher reproducibility relative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

Novel environmental application

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

* Presenting author

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



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

Ongoing work

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

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

100

GTL

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

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

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

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

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

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

* Presenting author

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

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

101

GTL

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

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

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

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

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

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

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

* Presenting author

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

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

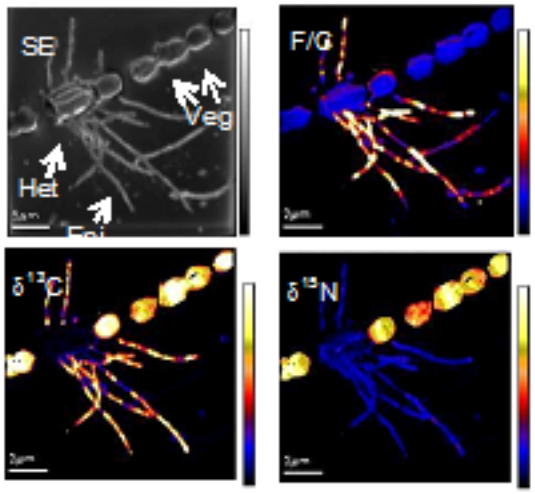


Figure. NanoSIMS images of a chain of 8 cells from the filamentous cyanobacterium *Anabaena* sp. SSM-00 with cells of the α-Proteobacterium epibiont *Rhizobium* sp. WH2K attached to a heterocyst. Images were taken after 24 h of incubation with $H^{13}CO_3^-$ and $^{15}N_2$. SE: Secondary electron image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

References

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

102

GTL

High Throughput Comprehensive and Quantitative Microbial and Community Proteomics

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

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

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

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

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

* Presenting author

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

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

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

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

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

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

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

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

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

* Presenting author

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

103

GTL

Proteomics Driven Analysis of Microbes and Microbial Communities

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

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

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

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

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

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

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

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

* Presenting author

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

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

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

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

Acknowledgements: This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute through Contract No. DE-AC05-76RLO 1830. Parts of this project is component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus.

104

GTL

Biofilm Growth Technologies for Systems Biology

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

J. Craig Venter Institute, La Jolla, California

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

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

* Presenting author

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

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

105

GTL

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

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

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

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

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

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

* Presenting author

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

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

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

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

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

106

GTL

Proteogenomics of Two Environmentally Relevant Microbial Communities

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

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

* Presenting author

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

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

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

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

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

* Presenting author

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

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

107

GTL

Structure and Function for Novel Proteins from an Extremophilic Iron Oxidizing Community

Korin Wheeler,[1]* Yongqin Jiao,[1] Steven Singer,[1] Adam Zemla,[1] Nathan VerBerkmoes,[2] Robert Hettich,[2] Daniela Goltsman,[3] Jill Banfield,[3] and **Michael Thelen**[1] (thelen1@llnl.gov)

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

Project Goals: With integrated metagenomic and proteomic datasets as a foundation, we are establishing a combination of computational and experimental methods to determine the structure and function of the several hundred proteins of unknown function within our well-defined, acid mine drainage model system. Our studies will enable a system-wide understanding of each protien and its role in cellular pathways and intracellular communication in this extremophilic microbial community.

As information from proteomic and genomic analyses rapidly escalates, the number of genes and proteins of unknown function continually expands. Yet methods to understanding these novel proteins, often key to understanding unique aspects of niche adaptation, are only just emerging. Because extreme environments are geochemically distinct and biologically limiting, low complexity ecosystems like that of acid mine drainage are ideal for such studies. Genomic and proteomic analysis of samples collected from the Richmond Mine in Iron Mountain (Redding, CA) have provided an initial survey of the genes and proteins that function within the community; however, it remains unclear how the numerous unique proteins facilitate survival under these conditions. With integrated metagenomic and proteomic datasets as a foundation[2], we are using a combination of computational and experimental methods to determine the structure and function of the several hundred proteins of unknown function within our model system.

Initial studies center on a high-throughput computational approach for predicting structure and function for 421 novel proteins from the dominant species in the community. We have developed a structural modeling system to compare these proteins to those of known structure (AS2TS)[2], resulting in the assignment of structures to 360 proteins (85%) and functional information for up to 75% of the modeled proteins. Detailed examination of the modeling results reveals the roles of many of the novel proteins within the microbial community. Protein classes (e.g., hydrolases, oxidoreductases) and families (metalloproteins, tetratricopeptide [TPR] repeats) that are highly represented in the community are now being targeted in experimental work. Complementing structure-function studies are biochemical and molecular biological techniques. Affinity chromatography has enabled enrichment of novel proteins with specific functions or active sites moieties. Environmental DNA clone libraries have facilitated screening of bacterial colonies for hydrolytic enzymes, including proteases, phosphatases, amylases, and lipases. Further to this molecular approach, a bacterial two-hybrid screen has been established to identify proteins interacting with our novel proteins, such as a novel iron oxidizing cytochrome and 27-repeat TPR protein.

References

1. Ram et al, 2005, *Science* 308:1915-20, "Community Proteomics of a Natural Microbial Biofilm."
2. Zemla et al, 2005, *Nucleic Acids Res* 33 (Web Server issue):W111-5, "AS2TS system for protein structure modeling and analysis."

This work was funded by the DOE Genomics:GTL Program and was performed under the auspices of the DOE by the University of California, Lawrence Livermore National Laboratory under contract W-7405-Eng-48.

* Presenting author

108

GTL

Purification and Characterization of Viruses From an Acid Mine Drainage System

Kimberly Pause[1]* (kpause@marine.usf.edu), Christine Sun,[2] Paul Wilmes,[2] Brett Baker,[2] Luis R. Comolli,[3] Chongle Pan,[4] Robert Hettich,[4] Nathan VerBerkmoes,[4] Jillian F. Banfield,[2] and **Mya Breitbart**[1]

[1]University of South Florida, Saint Petersburg, Florida; [2]University of California, Berkeley, California; [3]Lawrence Berkeley National Laboratory, Berkeley, California; and [4]Oak Ridge National Laboratory, Oak Ridge, Tennessee

Project Goals (Abstracts 109-111): The objective of this project is to develop integrated genomic and proteomic (proteogenomic) methods to study virus-microbial community interactions in bioreactor-grown and natural bacterial/archaeal biofilm communities. Simultaneous genomic analysis of microbes and viruses will be used to evaluate patterns of horizontal gene transfer and evolution. Since viruses tend to be host-specific predators, it is expected that they can drive shifts community structure by selecting for resistant strains. Although this phenomenon is common in laboratory chemostats, it has not been shown definitively for natural communities. Proteomics will be used to decipher metabolic interplay through monitoring of protein production before, during, and after viral infection, while molecular methods will be used to correlate these findings with changes in the community membership. We will also examine the hypothesis that CRISPR-associated Cas proteins, which may represent a component of a microbial immune system for viral defense, are highly produced in response to virus exposure and test the prediction that the genomes of immune strains encode small RNAs complementary to phage genes. Because viral predation can undermine microbially-based technologies such as microbial fuel cells, bioethanol production, and environmental remediation, methods to study microbe-virus interactions are widely relevant to DOE missions.

Viruses play important roles in biogeochemical cycling, horizontal gene transfer, and defining the community composition of their hosts. However, we are only beginning to understand the identity and diversity of viruses in the environment. Metagenomic sequencing has recently been used to examine viral communities from a variety of environments, including seawater, marine sediment, soil, and human feces. These studies have revealed a high degree of novelty and diversity amongst environmental viruses.

The extremely high diversity of viral communities has impeded studies of virus-host interactions in natural systems. The low microbial diversity of the acid mine drainage (AMD) community from Iron Mountain, CA provides an ideal setting for studying virus-host interactions. The bacterial and archaeal communities at the Iron Mountain AMD site have already been extensively characterized, leading to the identification of virus-derived spacer sequences in the CRISPR loci (see Sun et al. poster). An important next step is to examine purified virus particles from the AMD biofilm in order to compare CRISPR loci to coexisting viruses.

Here we describe methods for viral purification from an AMD biofilm, and an initial characterization of these viruses. A combination of filtration and density-dependent centrifugation was successfully used to purify virus particles from the AMD biofilm. Transmission electron microscopy of these viruses revealed icosahedral capids typical of bacteriophage, as well as unusual morphologies similar to archaeal viruses from hot springs and other extreme environments. In addition, 3-D reconstructions using cryo-electron tomography of lower abundance ultra-small archaea (ARMAN) from AMD biofilms revealed several cells under phage attack. All of the infected ARMAN cells have altered ultrastructure, not seen in uninfected cells, including very electron dense cytoplasmic proteins. Metagenomic sequencing of the viral nucleic acids is currently underway. Detailed analysis of the viruses from the AMD biofilm will allow us to determine if viral diversity correlates with host diversity, test "Kill-the-Winner" dynamics in a natural environment, further describe the role of CRISPRs in viral resistance, understand temporal and spatial variation of virus-host systems, and gain a deeper appreciation for phage and archaeal virus diversity and evolution.

This research was supported by a grant from the DOE Genomics:GTL program (2007).

* Presenting author

109

GTL

Development of Mass Spectrometry and Proteome Informatic Approaches to Analyze the Role of Virus–Microbe Interactions in Natural Microbial Communities

Nathan VerBerkmoes[1]* (verberkmoesn@ornl.gov), Alison Russell,[1] Chongle Pan,[1] Robert L. Hettich,[1] Manesh Shah,[1] Kim Pause,[2] Mya Breitbart,[2] Christine Sun,[3] Brian Thomas,[3] Paul Wilmes,[3] Anders Andersson,[3] and Jillian Banfield[3]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [2]University of South Florida, St. Petersburg, Florida; and [3]University of California, Berkeley, California

Project Goals: See goals for abstract 109.

Microorganisms comprise the majority of extant life forms and play key roles in a wide variety of health and environmental processes, yet little is known about the nature and driving forces of their diversification. Although the roles of viruses in microbial evolution are widely recognized, neither the details of viral-microbial interactions nor the impact of virus on microbial community structure are well understood. Community genomic and proteomic (proteogenomic) methods have been established for analyzing the roles and activities of uncultivated bacteria and archaea in natural multi-species consortia (Ram, *Science* 2005, Lo *Nature* 2007, Markert, *Science*, 2007). Notably lacking are methods for monitoring viral activity in communities, tracking virus predation, and determining the consequences of viral predation for ecosystem structure. This is broadly important because it is well established that viruses can control microbial abundances, influence microbial evolution, contribute to microbial adaptation by transferring metabolic genes, and have profound effects on carbon, nutrient, and metal cycling. Studies are mostly limited by the complexity of natural consortia, which makes it difficult to observe and correlate changes in virus and host community structure over time. Our work focuses on acid mine drainage (AMD) biofilms because of their relative simplicity and established utility as a model system for development of methods for cultivation-independent analyses. Our goal is to develop proteomics and informatic techniques to characterize the microbial response to viral attacks as well as to identify the viral proteins.

Our general method for proteome characterization of microbial communities has been well established and tested. It was originally developed on the model AMD system but has been extended to other microbial communities including waste water sludge samples, enriched microbial communities from oceans and ground waters, human gut microflora, and other complex systems. The proteomics pipeline involves non-biased cell lyses via chemical methods and/or sonication, protein denaturation and reduction, protein digestion via sequencing grade trypsin. The complex peptide mixture is then separated via two dimensional chromatography (SCX-RP) coupled on-line with rapid scanning electrospray mass spectrometers (LTQs and LTQ-Orbitraps). The peptide masses and MS/MS spectra (sequencing information) are compared to protein sequencing predicted from genomic sequences via search algorithms such as SEQUEST. Peptides are identified and then matched with protein(s) they originated from.

While these methods have been shown to work well with microbial communities, even if only moderately closely related reference genomes exist, there may be new challenges for identification of viral proteins. Genomic analyses that reconstructed population genomic datasets for five complex AMD virus populations that predate bacteria and archaea (as well as for a large plasmid/phage and for many partially sampled viral populations) revealed extraordinary levels of sequence diversity, as well as rapid shuffling of sequence motifs by recombination events. This will complicate proteomic analyses because enough peptides may differ from the sequenced peptides that protein identification will be precluded. In addition, many viral proteins may be present at abundance levels that are so low that they will be difficult to identify by standard methods.

Our primary objective is to develop new proteomic approaches to integrate analyses of virus-microorganism interactions into studies of the structure and dynamics of DOE-relevant microbial communities. The main challenge for using proteomics to study viruses and microbial-viral interactions is likely to be peptide (and thus protein) identification. As noted recently, database-searching programs (such as SEQUEST) identify peptides by "looking up the answer in the back of the book" (Sadygov et al. 2004; where the genome sequence is the answer section of the book). This approach has worked extremely well for genomically-characterized microbial isolates. It also works adequately for organisms whose sequences are close to the genomically characterized type because proteins can be identified using the subset of peptides that do not contain amino acid substitutions (Denef, *JPR* 2007). In cases where multiple candidate

* Presenting author

peptide sequences are available, we can distinguish the peptides and thus identify the protein variants (Lo et al., 2007). The largest challenge for proteomics of natural viral consortia is likely to arise when candidate virus peptide sequences are not available (i.e., the virus differs in sequence from any genomically characterized virus). Little is known about population level heterogeneity in viruses in the environment. Given their fast evolutionary rates, it is likely that pure database searching methods will be inadequate. As genomic re-sequencing of every sample is not practical, we will augment database searching with de novo sequencing (sequence tagging) methods.

The first method that we have adopted to deal with the challenge of virus sequence heterogeneity is to predict proteins in all reading frames from sequence reads that are assembled into composite virus genome fragments. This step will generate a much more extensive database of reference sequences for peptide and protein identification, with the added advantage that identified peptides will assist in identification of the correct ORF calls. Another component involves modification of protocols for enzymatic protein digestion so as to produce smaller peptides that will be less likely to differ from predicted protein sequences. Finally, and most importantly, we are investigating *de novo* sequencing programs that aim to derive complete or partial amino acid sequences from MS/MS scans without complete information from protein sequence databases. Although *de novo* sequencing approaches seem straightforward, they have not been widely applied in proteomics due to low data quality and software limitations. We have optimized methods for collection of high-resolution MS/MS scans with the LTQ-Orbitrap and are currently testing the existing *de novo* programs, including PEAKS and PepNovo (Frank, *Anal Chem* 2005) for their ability to correctly *de novo* sequence known peptides. Due to limitations of each of these programs we have designed an in house *de-novo* sequencing program specifically designed to use high resolution Orbitrap MS/MS data. Initial results indicate that the program has high accuracy. We are currently testing this program using microbial community proteome data from samples known to contain a bacterium with one of the two available genome sequences. To address the challenge of viral protein abundance, we are utilizing the well studied and understood *E. coli* / MS2 phage system. The goal of this study is to determine the effect of infection on the *E. coli* proteome and to determine the ability of proteomics to detect known viral protein sequences during the course of infection. In addition, we will characterize the effects of the MS2 phage on the *E. coli* proteome over a time series from initial infection to final death of the *E. coli* culture. Once we have resolved challenges associated with sequence variation and protein abundance levels, proteomics will be used to characterize natural biofilm communities, laboratory bioreactor communities, and laboratory co-cultures and microbial isolates. We will characterize these systems before, during and after viral infection and conduct time series experiments to monitor virus and microbial interactions and co-evolution.

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

110

GTL

Community Proteogenomic Studies of Virus–Microbe Interactions in Natural Systems

Christine Sun[1]* (christine_sun@berkeley.edu), Anders Andersson,[1] Paul Wilmes,[1] Mya Breitbart,[2] Chongle Pan,[3] Brian Thomas,[1] Nathan VerBerkmoes,[3] Robert Hettich,[3] and **Jillian F. Banfield**[1]

[1]University of California, Berkeley, California; [2]University of South Florida, Saint Petersburg, Florida; and [3]Oak Ridge National Laboratory, Oak Ridge, Tennessee

Project Goals: See goals for abstract 109.

Viruses (archaeal viruses and bacteriophages) play central roles in microbial evolution and, through predation of their microbial hosts, shape the composition and functionality of ecosystems. Yet, little knowledge exists about the extent of viral population diversity and the dynamics of virus-host interactions in natural communities. The recent discovery that genomic clustered regularly interspaced short palindromic repeat (CRISPR) loci of bacteria and archaea encode virus-derived spacer sequences that provide acquired viral resistance presents a unique opportunity to examine virus-host interactions. While most natural environments harbor diverse microbial and viral populations, a model natural microbial community—the low diversity, acid mine drainage (AMD) community from Iron Mountain, CA—provides the opportunity to comprehensively examine the dynamics between CRISPR regions and viruses.

CRISPR loci within the genomes of bacteria and archaea assembled from community genomic datasets for AMD biofilms typically demonstrate high levels of heterogeneity in terms of their spacer sequence complements, resulting from spacer gain and loss. Spacers extracted *in*

* Presenting author

silico from the CRISPR loci were used to identify viral sequences. Viral genomes assembled from spacer-bearing sequences exhibit evidence of very extensive recombination, with a level of sequence rearrangement sufficient to enable viruses to elude targeting by CRISPR-encoded spacer sequences. Viruses can be linked with their hosts through the CRISPR loci, making it possible to determine host ranges. Because CRISPRs incorporate new spacers with sequences that match coexisting viruses in a unidirectional manner, the loci provide a record of recent viral exposure.

A limitation for studies of the role of viruses in the ecology of natural communities is the under-sampling of certain virus types by the CRISPR loci and the limited number of CRISPR spacers that can be recovered for each organism. To address the first challenge, we have generated virus concentrates from natural samples and are currently characterizing these by electron microscopy and genomics (see USF poster). Secondly, to expand upon the set of CRISPR spacers available for identification of viral sequences in community genomic datasets, an extensive spacer database was created for the dominant AMD bacterial organism (*Leptospirillum* group II) via 454 FLX pyrosequencing. A large number of these spacers map to a reconstructed 56 kb genome of a virus (AMDV1) inferred to target *Leptospirillum* group II. Despite over 500,000 total spacers sequences (~17,000 of them unique), rarefaction curves did not saturate, indicating rapid and extensive divergence of CRISPR loci within natural populations.

The reconstruction of sequences from natural populations of viruses provides the opportunity to generate databases of encoded proteins. However, the challenges for proteomics are considerable. At this early stage in the project, we have focused on the development of methods for protein identification that address the challenge of virus-to-virus sequence divergence. Most promising from the proteomics standpoint are *de novo* peptide sequencing methods (see ORNL poster). In ongoing work, we will evaluate the ways in which the CRISPR machinery target and sample genomes of their viral predators, examine questions related to spatial and temporal variation in viral and microbial communities, and probe the physiological interactions that occur between bacteria, archaea, and virus populations, both in laboratory biofilm cultures and in the natural environment.

This research was supported by a grant from the DOE Genomics:GTL program (2007).

111

GTL

Molecular Signatures of the Past

Elijah Roberts[1]* (erobert3@scs.uicu.edu), Jonathan Montoya,[2] Anurag Sethi,[2] **Carl R. Woese**,[3] and Zaida Luthey-Schulten[1,2,3]

[1]Center for Biophysics and Computational Biology, [2]Department of Chemistry, and [3]Institute for Genomic Biology, University of Illinois, Urbana, Illinois

Project Goals: Ribosomal signatures, idiosyncrasies in the ribosomal RNA (rRNA) and/or proteins, are characteristic of the individual domains of life. From these studies, we propose the ribosomal signatures are remnants of an evolutionary phase transition that occurred as the cell lineages began to coalesce and should be correlated with signatures throughout the fabric of the cell and its genome.

Ribosomal signatures, idiosyncrasies in the ribosomal RNA (rRNA) and/or proteins, are characteristic of the individual domains of life. As such, insight into the evolution of the modern cell can be gained from a multi-dimensional comparative analysis of their manifestation in the translational apparatus. In this work, we identify signatures in both the sequence and structure of the rRNA, analyze their contributions to the signal of the universal phylogenetic tree using both sequence- and structure based methods, and correlate these RNA signatures to differences in the ribosomal proteins between the domains of life. Domain specific ribosomal proteins can be considered signatures in their own right and we present evidence that they evolved at the same time as the signatures in the ribosomal RNA and therefore should not be considered recent inventions. Furthermore, we demonstrate that signatures in the rRNA coevolved with the universal ribosomal protein S4. Given S4's role in the decoding center of the ribosome, this coevolution suggests a method by which evolution may use the ribosomal proteins to fine-tune translation in different environments. From these studies, we propose the ribosomal signatures are remnants of an evolutionary phase transition that occurred as the cell lineages began to coalesce and should be correlated with signatures throughout the fabric of the cell and its genome.

* Presenting author

112

GTL

Gene Synthesis by Circular Assembly Amplification

Duhee Bang* (dbang@genetics.med.harvard.edu) and **George M. Church**

Department of Genetics, Harvard Medical School, Boston, Massachusetts

We developed a novel gene-synthesis technology[1] to effectively reduce gene synthesis error rates by a factor of ten compared to conventional methods. Gene synthesis is playing an increasingly important role in biological research. Commonly employed methods however are highly prone to errors due to the errors originating in the synthetic oligonucleotides. In our approach, exonuclease-resistant circular DNA is first constructed via the simultaneous ligation of oligonucleotides. Subsequent exonuclease degradation of the resulting ligation mixture eliminates error-rich linear products, thereby significantly improving gene-synthesis quality. By combining circular assembly amplification with the use of mismatch cleaving endonucleases we have achieved error rates of 0.025%. The method has been used to construct genes encoding Dpo4, a small thermo-stable DNA polymerase, and to construct highly repetitive DNA sequences which are not amenable to traditional synthesis methods. By adapting a uridine-cleavage strategy, we have also used the circular assembly amplification to synthesize large (>4 kb) constructs. This method promises to significantly cut the cost of gene synthesis, as the assembly of ~1kb gene (an average length of a gene) can be done in single cycle with a smaller amount of sequencing required to find a perfect construct.

Reference

1. Bang D, Church GM. Gene synthesis by circular assembly amplification. Nat Methods. 2007 Nov 25 (published online).

113

GTL

Mycoplasma Genome Synthesis and Transplantation: Progress on Constructing a Synthetic Cell

Daniel G. Gibson* (dgibson@jcvi.org), Carole Lartigue, Gwynedd A. Benders, John I. Glass, Clyde A. Hutchison III, Hamilton O. Smith, and **J. Craig Venter**

The J. Craig Venter Institute, Rockville, Maryland

Project Goals: Synthetic cell production.

Mycoplasma genitalium is an approximately 300nm diameter wall-less bacterium that has the smallest known genome of any cell that can be grown in pure culture in the laboratory. When grown under ideal conditions in a rich, serum-containing medium, as many as 100 genes appear to be dispensable based on one-gene-at-a-time transposon mutagenesis. In order to better understand the essence of a minimal cell, we are employing a synthetic genomics approach to construct a 582,970 bp *M. genitalium* genome. The synthetic genome will contain all the genes of wild type *M. genitalium* G37 except MG408, which will be disrupted by an antibiotic resistance marker to block pathogenicity and to allow for selection. Overlapping "cassettes" of 5-7 kb, assembled from chemically synthesized oligonucleotides, are being joined by *in vitro* recombination to produce intermediate assemblies of approximately 24 kb, 72 kb ("1/8 genome"), and 144 kb ("1/4 genome") and cloned as bacterial artificial chromosomes (BACs) in *Escherichia coli*. Once assemblies of all four 1/4 genomes are identified, the complete synthetic genome will be assembled and cloned in the yeast *Saccharomyces cerevisiae*. Minimization of the synthetic genome can be carried out by assembly of cassettes with individual genes deleted or by genome reduction with recombineering methods. Both approaches require the development of methods to transplant the synthesized genome into a receptive cytoplasm such that the donor genome becomes installed as the new operating system of the cell. As a step toward propagation of synthetic genomes, we completely replaced the genome of *M. capricolum* with one from *M. mycoides* LC by transplanting a whole genome as naked DNA. These cells that result from genome transplantation are phenotypically identical to the *M. mycoides* LC donor strain as judged by several criteria. The methods described are fundamental to the full development of synthetic biology.

* Presenting author

Molecular Interactions and Protein Complexes

114

GTL

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

Angeli Lal Menon[1]* (almenon@uga.edu), Aleks Cvetkovic,[1] Sarat Shanmukh,[1] Farris L. Poole II,[1] Joseph Scott,[1] Ewa Kalisiak,[2] Sunia Trauger,[2] Gary Siuzdak,[2] and **Michael W.W. Adams**[1]

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

Project Goals: The genes that encode multiprotein complexes (PCs) or post-translationally modified proteins (MPs), such as those that contain metal cofactors, in any organism are largely unknown. We are using non-denaturing separation techniques coupled to mass spectrometry (MS) analyses to identify PCs and MPs in the native biomass of *P. furiosus*. By analyzing the native proteome at temperatures close to 100°C below the optimum physiological temperature, we will trap reversible and dynamic complexes thereby enabling their identification and purification. Samples of the more abundant PCs and MPs obtained from native biomass are structurally characterized. This research is part of the MAGGIE project, the overall goal of which is to determine how Molecular Assemblies, Genes, and Genomics are Integrated Efficiently. The MAGGIE project is funded by the Genomics:GTL program of DOE with John Tainer, Scripps/LBL, as the PI.

http://masspec.scripps.edu/maggie/bacteria.php

Most biological processes are carried out by dynamic molecular assemblies or protein complexes (PCs), many of which include modified proteins (MPs) containing organic and/or inorganic cofactors. The composition and the protein components of these complexes are largely unknown. They cannot be predicted solely from bioinformatics analyses, nor are there well defined techniques currently available to unequivocally identify PCs or MPs. Directly determining the identity of PCs and MPs in native biomass can resolve some of these issues. We are currently using *Pyrococcus furiosus*, a hyperthermophilic archaeon that grows optimally at 100°C, as the model organism. Fractionation of native biomass close to 80°C below the optimal growth temperature using non-denaturing, chromatography techniques, should enable purification of both stable and dynamic PCs and MPs for further characterization.

Large scale fractionation of native *P. furiosus* biomass was carried out under anaerobic and reducing conditions. Cytoplasmic proteins were fractionated by ion-exchange chromatography generating 126 fractions. These were combined into fifteen pools and subjected to 25 additional non-denaturing, chromatographic columns generating a total of 1276 fractions from all 26 chromatography steps. The identification of potential native PCs and MPs was accomplished by analyzing fractions using native and denaturing PAGE, mass spectrometry (MS), bioinformatics and for protein and metal concentrations. To date, 967 distinct *P. furiosus* proteins have been identified from the cytoplasmic protein fractionation. Based on the co-elution of proteins encoded by adjacent genes (same DNA strand), 243 proteins are proposed to be contained in 106 potential heteromeric PCs. In addition, proteins in highly-salt washed membranes were detergent-solubilzed and further fractionated using ion-exchange chromatography. A total of 205 proteins were identified from bands on PAGE. Half of them contained potential transmembrane domains, and 77 are proposed to form 25 potential PCs, which include a complex of hydrogenase (12 subunits) and ATP synthase (9 subunits).

Chromotographic fractions were also analyzed by ICP-MS for 54 elements. A subset of the elements that were unambiguously identified in the first column fractions were selected for preliminary ICP-MS analysis of the fractions from the 15 second level chromatography columns. As examples, 12 Fe-, 12 Co- and 5 Zn-containing peaks from the first ion-exchange column were resolved into 48 Fe-, 39 Co- and 13 Zn-containing peaks, respectively, on the second level columns. Using criteria such as coincident protein LC/MS/MS profiles, the presence of putative metal or metal-cofactor binding motifs and homology to characterized metalloproteins, several metal peaks in these element profiles have been assigned to known metalloproteins, while others have been identified as being potential metalloproteins. As examples, 2 Fe- and 2 Co-containing known metalloproteins were identified in the first column fractions, while at least 2 Fe-, 7 Co- and 4 Zn-containing peaks were assigned to as yet uncharacterized proteins, which are potentially

* Presenting author

novel metalloproteins. In addition, a large number of metal peaks were observed that could not be assigned to any of the proteins identified by LC/MS/MS analyses.

The results of further characterization of the soluble and membrane-bound PCs identified in this study, and particularly those containing metals, will be discussed.

115

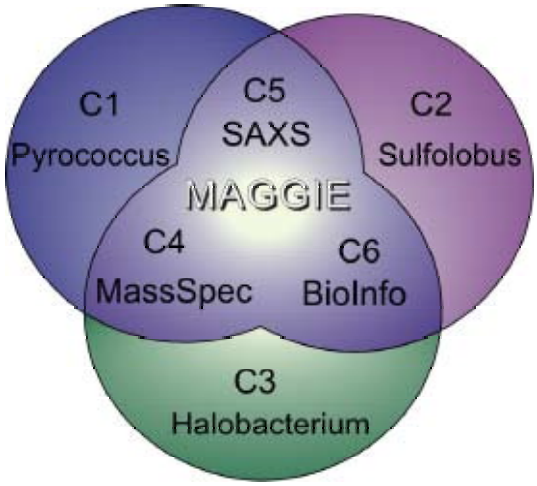
GTL

Molecular Assemblies, Genes, and Genomics Integrated Efficiently: MAGGIE

John A. Tainer[1,3]* (jat@scripps.edu), Greg L. Hura[1], Steven M. Yannone[1], Stephen R. Holbrook[1], Jane Tanamachi[1], Mike Adams[2], Gary Siuzdak,[3] and Nitin S. Baliga[4]

[1]Life Science and Physical Biosciences Divisions, Lawrence Berkeley National Laboratory, Berkeley, California; [2]University of Georgia, Athens, Georgia; [3]The Scripps Research Institute, La Jolla, California; and [4]Institute for Systems Biology, Seattle, Washington

Project Goals: MAGGIE (Molecular Assemblies, Genes, and Genomics Integrated Efficiently) will provide robust GTL technologies and comprehensive characterizations to efficiently couple gene sequences and genomic analyses with protein interactions and thereby elucidate functional relationships and pathways. To accomplish its goals, MAGGIE integrates an interdisciplinary team at Lawrence Berkeley National Lab with researchers at The Scripps Research Institute, the University of Georgia, the University of California Berkeley, and the Institute for Systems Biology into a unified Genomics:GTL program. MAGGIE thus focuses on providing an integrated, multi-disciplinary program and synchrotron facilities at the Advanced Light Source (ALS) to achieve efficient key technologies and databases for the molecular-level understanding of the dynamic macromolecular machines that underlie all of microbial cell biology. Three overall goals are 1) to facilitate instrument and technology development and optimizations though cross-disciplinary collaborations, 2) to comprehensively characterize complex molecular machines including protein complexes (PCs) and modified proteins (MPs) and 3) to provide critical enabling technologies and a prototypical map of PCs and MPs for the GTL Program. In concert, MAGGIE investigators will help to characterize microbial metabolic modularity and to provide the informed basis to design functional islands suitable to transform microbes for specific DOE missions.



MAGGIE integrates teams at Lawrence Berkeley National Lab and the Advanced Light Source (ALS) with researchers at the Scripps Research Institute, the University of Georgia, the University of California Berkeley, and the Institute for Systems Biology to achieve a molecular-level understanding of the dynamic macromolecular machines that underlie all of microbial cell biology. MAGGIE is providing improved technologies and comprehensive characterizations to efficiently couple gene sequences and genomic analyses with protein interactions and thereby elucidate functional relationships and pathways. The operational principle guiding MAGGIE objectives can be succinctly stated: protein functional relationships can be characterized as interaction mosaics that self-assemble from independent protein pieces and that are tuned by modifications and metabolites.

MAGGIE builds strong synergies among the program components to address long term and immediate GTL objectives by combining the advantages of specific microbial systems with those of advanced technologies. A key objective for the MAGGIE Program is therefore to characterize the Protein Complexes (PCs) and Modified Proteins (MPs) underlying microbial cell biology including responses to environment. A compelling overall goal is to help reduce the complexity of protein interactions to interpretable patterns though an interplay among experimental efforts of MAGGIE program members in molecular biology, biochemistry, biophysics, mathematics, computational science, and informatics.

MAGGIE investigators are working in concert to address GTL missions by accomplishing three specific goals: 1) develop and apply advanced mass spectroscopy and Small Angle X-ray Scattering (SAXS) technologies

* Presenting author

for high throughput characterizations of complex molecular machines including PCs and MPs, 2) create and test efficient mathematical and computational systems biology descriptions for protein functional interactions, and 3) provide both critical enabling technologies and a prototypical map of PCs and MPs for the GTL Program. Emerging databases are linked via the open-source Gaggle software system, which provides the efficient, flexible technology for communications across databases: http://gaggle.systemsbiology.org/docs/. Our new SAXS methodological treatise characterizes data interpretation tools to examine molecular interactions, flexibility, and conformational changes in solution relevant to understanding and predictions: http://bl1231.als.lbl.gov/2007/10/review_of_saxs_combined_with_c.php. The MAGGIE results are generally accessible on our website: http://masspec.scripps.edu/MAGGIE/index.php.

116

GTL

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

Stephanie Patterson,[1] Jill Fuss,[1] Kenneth Stedman,[2] Michael W.W. Adams,[3] Gary Siuzdak,[4] Trent Northen,[4] Ewa Kalisiak,[4] Sunia Trauger,[4] Nitin S. Baliga,[5] Stephen R. Holbrook,[1] John A. Tainer,[1,6] and **Steven M. Yannone**[1]* (SMYannone@lbl.gov)

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

Project Goals: As part of MAGGIE, we are developing high throughput recombinant DNA and native biomass technologies for the extremophilic organism *Sulfolobus solfataricus* which has a growth optimum at 80° C and pH 3.0. We are also exploiting the unique characteristics of Archaeal membranes to isolate membrane-protein as well as soluble protein complexes from native biomass. We have developed universally applicable approaches for enriching native biomass for protein complexes and established simplified proteomic analyses resulting in greater than 50% coverage of the *Sulfolobus* proteome with relatively simple and rapid procedures. We are developing the computational tools necessary to integrate protein fractionation, predicted/observed molecular mass, genomic organization, and co-fractionation data sets to identify novel protein complexes. Using MS/MS-based metabolomic analyses, we have identified an un-annotated trehalose synthesis pathway in *Sulfolobus* and are expanding this approach to validate homology-based functional predictions. We have begun the structural characterization of recombinant *Sulfolobus* proteins with small angle x-ray scattering (SAXS) at the advanced light source at LBNL, and will discuss the biological implications of these studies. We are testing the idea that the hyperthermophilic nature of *Sulfolobus* will allow us to "thermally trap" protein complexes assembled at 80° C by isolating these complexes at room temperature.

Dynamic protein-protein interactions are fundamental to most biological processes and essential for maintaining homeostasis within all living organisms. Understanding the networks of these protein interactions is of critical importance to understanding the complexities of biological systems. The MAGGIE project was conceived, in part, as a response to the DOE GTL initiative to develop technologies to map the proteomes of model organisms. In this project we are exploiting unique characteristics of members of extremophilic Archaea to identify, isolate, and characterize multi-protein molecular machines. We have teamed expertise in mass spectrometry, systems biology, structural biology, biochemistry, and molecular biology to approach the challenges of mapping relatively simple proteomes.

As part of MAGGIE, we are developing high throughput recombinant DNA technologies for the extremophilic organism *Sulfolobus solfataricus* which has a growth optimum at 80°C and pH 3.0. We are using a naturally occurring viral pathogen of this organism to engineer shuttle vectors designed for recombinant protein tagging and expression in the native *Sulfolobus* background. We are also exploiting the unique characteristics of *Archaeal* membranes to isolate membrane-protein as well as soluble protein complexes from native biomass. We have developed universally applicable approaches for enriching native biomass for protein complexes and established simplified proteomic analyses resulting in greater than 50% coverage of the *Sulfolobus* proteome with relatively simple and rapid procedures. We are developing the computational tools necessary to integrate protein frac-

* Presenting author

tionation, predicted/observed molecular mass, genomic organization, and co-fractionation data sets to identify novel protein complexes. Using MS/MS-based metabolomic analyses, we have identified an un-annotated trehalose synthesis pathway in *Sulfolobus* and are expanding this approach to validate homology-based functional predictions. We have begun the structural characterization of recombinant *Sulfolobus* proteins with small angle x-ray scattering (SAXS) at the advanced light source at LBNL, and will discuss the biological implications of these studies. We are testing the idea that the hyperthermophilic nature of *Sulfolobus* will allow us to "thermally trap" protein complexes assembled at 80°C by isolating these complexes at room temperature. Ultimately, we aim to identify metabolic modules suitable to transfer specific metabolic processes between microbes to address specific DOE missions while developing generally applicable molecular and biophysical technologies for GTL.

117

GTL

Metabolomic Profiling of a Hyperthermophile and the Characterization of Metabolite-Protein Interactions

Sunia A. Trauger[1]* (strauger@scripps.edu), Ewa Kalisak,[1] Jaroslaw Kalisiak,[1] Hirotoshi Morit,[1] Michael V. Weinberg,[2] Angeli Lal Menon,[2] Farris L. Poole II,[2] Michael W.W. Adams,[2] and Gary Siuzdak[1]

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

We have performed a comprehensive characterization of global molecular changes using the hyperthermophilic archaeon, *Pyrococcus furiosus,* as a model organism and using transcriptomic (DNA microarray), proteomic and metabolomic analysis as it undergoes a cold adaptation response from its optimal 95°C to 72°C. Metabolic profiling on the same set of samples show the down-regulation of many metabolites. However, some metabolites are found to be strongly up-regulated. An approach using accurate mass, isotopic pattern, database searching and retention time is used to putatively identify several metabolites of interest. Many of the up-regulated metabolites are part of an alternative polyamine biosynthesis pathway previously established in a thermophilic bacterium *Thermus thermophilus*. [1]

Arginine, agmatine, spermidine and branched polyamines N^4-aminopropylspermidine and N^4-(*N*-acetylaminopropyl)spermidine were unambiguously identified based on their accurate mass, isotopic pattern and matching of MS/MS data acquired under identical conditions for the natural metabolite and a high purity standard. For the branched polyamines N^4-aminopropylspermidine and N^4-(*N*-acetylaminopropyl)spermidine, both DNA microarray and semi-quantitative proteomic analysis using a label-free spectral counting approach indicate the down-regulation of a large majority of genes with diverse predicted functions related to growth such as transcription, amino acid biosynthesis and translation. Some genes are however, found to be up-regulated through the measurement of their relative mRNA and protein levels. A novel approach using metabolite immobilization for protein capture followed by proteomic analysis is used for the identification of protein partners which may interact with three polyamines. These were agmatine, spermidine and the novel metabolite N^4-(*N*-acetylaminopropylsprmidine) involved in the alternative polyamine biosynthetic pathway. Proteins identified using this method included unique proteins, as well as ones which were common to all three polyamines. Proteins identified using immobilized spermidine as bait, included SAM decarboxylase SpeD (PF1930) and S-adenosylmethionine synthetase (PF1866) which are directly involved in its probable biosynthetic pathway. Other proteins identified with spermidine immobilization are involved in translation such as PF1375 (translation elongation faction ef1), PF1367 (LSU ribosomal protein L7AE), PF1264 (translation elongation factor eIF-5a). Polyamines such as spermidine are known to play a critical role in hyperthermophiles in translation, ribosomal assembly and protein elongation and their interaction may reflect the specific affinity of spermidine for these proteins. For the novel metabolite N^4-(*N*-acetylaminopropylsprmidine), for which the enzymatic pathways involved in its synthesis remain unknown, a conserved hypothetical protein PF0607 was uniquely identified. While, proteins identified include those which are clearly interacting with other proteins and DNA that bind the immobilized metabolite, we believe this is a promising technique which could be used as an initial screen for uncovering enzymatic processes that underlie the biosynthesis of newly identified metabolites. The complimentary information obtained by the various 'omics' techniques are used to catalogue and correlate the overall molecular changes.

Reference

1. Ohnuma, M.; Terui, Y.; Tamakoshi, M.; Mitome, H.; Niitsu, M.; Samejima, K.; Kawashima, E.; Oshima, T., N1-aminopropylagmatine, a new

* Presenting author

polyamine produced as a key intermediate in polyamine biosynthesis of an extreme thermophile, Thermus thermophilus. *J Biol Chem* **2005,** 280, (34), 30073-82.

118

GTL

Protein Complex Analysis Project (PCAP): Project Overview

Lauren Camp,[1] Swapnil Chhabra,[1] Dwayne Elias,[3] Jil T. Geller,[1] Hoi-Ying Holman,[1] Dominique Joyner,[1] Jay Keasling,[1,2] Aindrila Mukhopadhyay,[1] Mary Singer,[1] Tamas Torok,[1] Judy Wall,[3] **Terry C. Hazen**,[1*] Simon Allen,[4] Gareth Butland,[1] Megan Choi,[1] Ming Dong,[1] Barbara Gold,[1] Steven C. Hall,[4] Bing K. Jap,[1] Jian Jin,[1] Susan J. Fisher,[4] Haichuan Liu,[4] Ramadevi Prathapam,[1] Evelin Szakal,[4] Peter J. Walian,[1] H. Ewa Witkowska,[4] Lee Yang,[1] Wenhong Yang,[1] **Mark D. Biggin**[1*] (mdbiggin@lbl.gov), Pablo Arbelaez,[1] Manfred Auer,[1] David Ball,[1] Myhanh Duong,[1] Robert M. Glaeser,[1] Bong-Gyoon Han,[1] Danielle Jorgens,[1] Jitendra Malik,[2] Hildur Palsdottir,[1] Jonathan P. Remis,[1] Dieter Typke,[1] **Kenneth H. Downing**,[1*] Steven S. Andrews,[1] Adam P. Arkin,[1,2] Steven E. Brenner,[1,2] Y. Wayne Huang,[1] Keith Keller,[1] Ralph Santos,[1] Max Shatsky,[1,2] and **John-Marc Chandonia**[1*]

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

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

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

Essential to this endeavor is the development of automated high throughput methods that are robust and allow for the comprehensive analysis of many protein complexes. Biochemical purification of endogenous complexes and identification by mass spectrometry is being coupled with in vitro and in vivo EM molecular imaging methods. Because no single method can isolate all complexes, we are developing two protein purification pipelines, one the current standard Tandem Affinity Purification approach, the other a novel tagless strategy. Specific variants of each of these are being developed for water soluble and membrane proteins. Our Bioinstrumentation group is developing highly parallel micro-scale protein purification and protein sample preparation platforms, and mass spectrometry data analysis is being automated to allow the throughput required. The stoichiometries of the purified complexes are being determined and the quaternary structures of complexes larger than 250 kDa are being solved by single particle EM. We are developing EM tomography approaches to examine whole cells and sectioned, stained material to detect complexes in cells and determine their localization and structures. New image analysis methods will be applied to speed

* Presenting author

determination of quaternary structures from EM data. Once key components in the interaction network are defined, to test and validate our pathway models, mutant strains not expressing these genes will be assayed for their ability to survive and respond to stress and for their capacity for bioreduction of DOE important metals and radionuclides.

Our progress during the second year of the project includes establishing both Gateway and Recombineering based pipelines for constructing genetically altered *D. vulgaris* strains; ramping up biomass production; establishing an optimized four-step tagless fractionation series for the purification of water soluble protein complexes from 400L of culture; establishing proof of principle data for the effectiveness of the tagless identification of protein complexes by mass spectrometry; scaling up tagless purification of inner and outer membrane complexes; identifying over 50 water soluble and membrane complexes; automating many aspects of mass spectrometry data analysis; establishing a TAP pipeline; determining the structure of five additional complexes by single particle EM; developing an improved automated particle picking method for EM images of purified complexes; establishing fluorescent SNAP tag labeling of complexes in biofilms; discovering differences in activity between cells associated with fibers in biofilms; and establishing a novel approach (WIST) for automated construction of database web interfaces that speeds database construction and using it to build LIMS modules to store data from several parts of our workflow. Further details on these and other results are provided in the Subproject specific posters.

119

GTL

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

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

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

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

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

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

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

* Presenting author

Tagless purification of water soluble complexes. We have developed an optimized four-step fractionation scheme for the tagless purification strategy that uses protein from 400L of culture and have used it to identify and purify over 50 homomeric and heteromeric water soluble protein complexes from just 0.2% of the fraction space. We have established an efficient, highly reproducible mass spectrometry sample preparation protocol that uses 96-well PVDF multiscreen plates and is effective with the iTRAQ methodology we have adopted to quantitate the relative abundances of polypeptides in different chromatographic fractions. Methods for preparing protein samples suitable for single particle EM analysis are being refined, including the use of different crosslinking reagents to stabilize complexes on EM grids. To date, 16 complexes have been sent to the EM Subproject for structural determination. As a result, structures at various resolution have been obtained for Pyruvate Ferredoxin Oxidoreductase, GroEL, a putative protein DVU0671, PEP synthase, and 6,7-dimethyl-8-ribityllumazine synthase.

Tagless purification of membrane complexes. Over the past year we have isolated membrane protein complexes using developmental protocols featuring several chromatographic steps (ion exchange, hydroxyapatite and molecular sieve) and blue native gel electrophoresis. For this work, mild detergents have been used to sequentially solublize proteins of the inner and outer membrane. With this approach and an improved protocol for preparing mass spectrometry samples, 20 membrane protein complexes (homomeric and heteromeric) have been identified. Membrane protein complex samples have also been prepared for preliminary electron microscopy analysis and delivered to the EM group.

Tandem Affinity Purification of water soluble complexes. We have completed trials of different TAP tag combinations for protein complex purification from *D. vulgaris*. Initial tests have compared the efficiency of the Sequential Peptide Affinity (SPA) tag and the Strep-TEV-FLAG (STF) tag. We have confirmed that both tags can purify proteins synthesized in *D. vulgaris* with comparable high yield and low background binding properties. We have also completed optimization trials to determine the quantities of *D. vulgaris* biomass required for purification of protein complexes in amounts sufficient for identification by mass spectrometry. Strains bearing individually tagged genes will be generated for this high throughput purification pipeline using high-throughput cloning strategies currently being deployed to construct tagged *D. vulgaris* genes rapidly and efficiently in *E. coli*.

Automation of protein complex purification. We have developed a multi-channel, native gel electrophoresis instrument for high resolution protein separation and automated band collection. This instrument can separate samples of protein mixtures from ~20Kd to ~600Kd and elute a protein band into a 200 μl fraction, without noticeable loss of sample. The use of this free-flow electrophoresis apparatus will greatly assist our efforts to achieve high throughput and provide an additional means of obtaining specimens in amounts appropriate for EM studies.

Mass spectrometry. We have worked in parallel both to further optimize mass spectrometry sample preparation and data acquisition and to discover *DvH* complexes separated *via* a tagless protein complex identification pipeline. We have implemented use of an internal protein standard to monitor recoveries during sample preparation of iTRAQ-labeled tryptic peptides, allowing us to normalize quantitation results. We have analyzed recoveries of peptides varying in hydrophobicity, charge and size from the PVDF membranes used in our high throughput sample preparation method. We have also demonstrated that our mass spectrometry protocols are accurate for samples of lower concentration and lower total protein load, down to 2 μg. Protein complex discovery was performed on 0.2% of the fraction space of soluble protein derived from 400 L prep (76 sizing column fractions). *In toto*, 160 polypeptides were matched to 2 or more peptides and a further 70 polypeptides matched to a single peptide. At least 7 heteromeric complexes orthologous to *E. coli* complexes, 3 novel heteromeric complexes, and 39 known homomeric complexes were identified as well as tens of additional polypeptides whose size migration suggested are part of a complex. Three of the heteromeric complexes were also examined by TAP, giving similar results to those obtained by the Tagless approach. The large number of complexes detected within a very small portion of the overall analytical space indicates that the tagless strategy holds high potential for characterizing the bacterial interactome. We have also employed Synapt HD mass spectrometer (Waters) to characterize heterogeneity within dissimilatory sulfite reductase and found that discrete forms of the complex differ in subunit stoichiometry.

Integrated mass spectrometry data acquisition and automated data analysis. To handle the mass spectrometry data for the large number of protein fractions generated by our tagless strategy, an integrated data acquisition and automated processing pipeline has been developed. This integrates commercial ProteinPilot software with several home-developed processing tools. ProteinPilot fetches data directly from the Oracle database of our

* Presenting author

AB4800 mass spec. instrument and produces lists of identified proteins and their relative concentrations. The tools we have developed generate normalized elution profiles for each detected protein and allow automatic initiation and monitoring of the protein identification process once MS/MS data are available on the Oracle database. To reduce redundant MS/MS data acquisitions and improve coverage of less abundant proteins, an iterative and intelligent data acquisition has been integrated into the pipeline. We have developed a comparator that pulls data and information from various modules in the pipeline and generates inclusion/exclusion lists to direct subsequent data acquisition. To complete the system, a dynamic instrument command control module is being developed that takes inclusion/exclusion lists and predictions of peptide retention times as inputs and produces more effective MS/MS data acquisition requests. In operation it will control and monitor AB4800's data acquisition and provide on-the-fly assessments of spectrum quality and peptide id feasibility. A clustering analysis tool has also been developed that can effectively identify protein complexes. This automatic tool takes protein elution profiles as input, performs several preprocesses including peak identification, de-noising and peak-overlap characterization, and then calculates correlation coefficients of every pair of peaks and ranks and sorts results to identify components of protein complexes.

120

GTL

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

Adam P. Arkin,[1,2] Steven E. Brenner,[1,2] Max Shatsky,[1,2] Ralph Santos,[1] Wayne Huang,[1] Keith Keller,[1] and **John-Marc Chandonia**[1,2]* (jmchandonia@lbl.gov)

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

Project Goals: The Data Management and Bioinformatics component of the Protein Complex Analysis Project (PCAP) has two major goals: 1. to develop an information management infrastructure that is integrated with databases used by other projects within the Virtual Institute for Microbial Stress and Survival (VIMSS), and 2. to analyze data produced by the other PCAP subprojects together with other information from VIMSS to model stress responses relevant to the use of *D. vulgaris* and similar bacteria for bioremediation of metal and radionuclide contaminated sites. In addition to storing experimental data produced by the PCAP project, we will assess the quality and consistency of the data, and compare our results to other public databases of protein complexes, pathways, and regulatory networks. We will prioritize proteins for tagging, TAP, and study by EM based on analysis of VIMSS data and other bioinformatic predictions. All data we obtain on protein interactions will be analyzed in the context of the data currently stored in VIMSS. One of the primary goals of VIMSS is the creation of models of the stress and metal reduction pathways of environmental microbes. Ultimately, we wish to analyze PCAP data in such a way as to automatically generate hypothetical models of cellular pathways, which will be validated by comparison to experimental observations.

We are developing a modular LIMS system to store data and metadata from the high-throughput experiments undertaken by the other PCAP subprojects. Each module of the LIMS corresponds to a step in the experimental pipeline. We have developed WIST (Workflow Information Storage Toolkit), a template-based toolkit to facilitate rapid LIMS development. WIST allows LIMS programmers to design multi-step workflows using modular core components, which can be added and arranged through a simple, intuitive configuration and template mechanism. WIST uses the templates to create unified, web-based interfaces for data entry, browsing, and editing. We have deployed WIST in an updated version of the tagless purification module, the tagged purification module, and as a component of our automated pipeline for sequence validation of high throughput constructs.

We have also prioritized proteins for tagging, TAP, and study by electron microscopy based on analysis of gene expression data from the VIMSS Environmental Stress Pathway Project (ESPP) and bioinformatic predictions. To date, we have identified 1217 *D. vulgaris* proteins as high-priority targets for tagging by the PCAP Microbiology Core. 265 of these proteins have already been identified as likely components of multimeric complexes by mass spectroscopy of fractions purified using the tagless pipeline. We will cross-validate our results by comparing the composition of complexes characterized by both the tagless and tagged purification pipelines. In addition, *D. vulgaris* orthologs of proteins belonging to previously

* Presenting author

characterized complexes from other organisms have been selected as tagging targets. This was done in order to study the degree to which stable inter-protein interactions are conserved between orthologs, and to establish a baseline characterization of potential complexes to compare with the same proteins under stress conditions. We also plan to study the degree to which correlated expression in microarray experiments may be used to predict stable protein interactions.

121

GTL

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

Manfred Auer, David Ball, Myhanh Duong, Danielle Jorgens, Hildur Palsdottir, Jonathan Remis, and **Kenneth H. Downing*** (KHDowning@lbl.gov)

Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California

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

The broad aim of this Subproject of PCAP is to demonstrate the feasibility of using electron microscopy for high-throughput structural characterization of multi-protein complexes in microbes of interest to DOE. Our goal is to determine the spatial organization and relative locations of large multi-protein complexes within individual, intact microbes, as well as in microbial communities, using cryo-EM tomography and novel tag-based labeling approaches.

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

We also employ plastic-section electron microscopy to study both planktonic and biofilm forms of microbes of interest. This approach has the advantage that it lends itself more easily to labeling – and thus localizing – genetically tagged proteins. Sectioning is also the only technique that can provide images of specimens that are too thick to image as whole-mount materials, while still retaining nanometer resolution. The ultimate goal in using plastic-section microscopy is thus to provide the most complete and accurate information possible about the status of multi-protein complexes, and to do so in a way that can then be used to improve mathematical modeling of cellular responses under the environmental conditions that require bioremediation.

In order to take advantage of the genetic tools that allow tagging of specific proteins for localization by both light and electron microscopy, we are focusing on several fluorescent reagents that can be characterized in the light microscope and then photoconverted to electron-dense signals for electron microscopy. This is quite a new endeavor for anaerobic bacteria such as *D. vulgaris*, which produce high levels of H_2S. Our initial focus is on morphology of biofilms in which we see a number of structures that have yet to be characterized in *D. vulgaris*. We grow biofilms of *D. vulgaris* on cellulose dialysis tubing or sheets, where the biofilms cover almost the entire available surface area. Samples are high pressure frozen and freeze-substituted in order to optimize preservation of structural details. Electron microscopic analysis of biofilm sections reveals loose packing of *D. vulgaris* within the biofilm EPS. Interestingly we found filamentous string-like metal precipitates near the *D. vulgaris*, which may point to structures not unlike the well-characterized *Shewanella* nanowires, which are known to be instrumental in extracellular metal reduction. Variations in the deposition patterns indicate that metal reduction activity

* Presenting author

varies between neighboring cells in biofilms. Using microwave processing of 6-day old mature *D. vulgaris* biofilm, we have confirmed the existence of such metal strings that extend dozens and possibly hundreds of microns. Interestingly, we found that intact cells are associated with these metal strings, whereas areas devoid of such metal strings only contain cell debris, suggesting that these metals strings contribute to cell survival in such stationary biofilms.

We have developed on-grid culturing methods for rapid study of such features in cell monolayers grown under various environmental conditions, and found the presence of filamentous structures of ~ 7nm in diameter that were associated with metals precipitated out of a Uranium-containing solution.

We have tested ReAsH and SNAP-labeling of several strains of *D. vulgaris* in which proteins have been tagged by members of the PCAP Microbiology group. The SNAP-labeling appears promising as judged by light microscopy. In-invo and in-vitro labeling of tagged proteins, before and after cell lysis, respectively, followed by SDS PAGE suggests specific binding for the SNAP-tag reagent. We found large variations in the labeling intensity of planktonic cells, with only about 20-30% of these genetically identical bacteria displaying strong labeling. We have ruled out the possibility that this difference is due to variability of reagent access or vitality of the cells, suggesting that there are large differences in protein expression levels even at the planktonic state. We speculate that differences in protein expression levels may be the reason for cell-to-cell differences of metal reduction capability as seen in planktonic cells and in biofilms. We are currently optimizing the photoconversion of the fluorescence signal both for planktonic cells and biofilms.

While the intact *D. vulgaris* cells are generally thicker than optimal for high resolution electron tomography, initial results show that various approaches to specimen preparation can produce samples in which a wealth of internal detail can be visualized. We have begun to characterize the initial morphological responses to oxidative stress, which are particularly dramatic in cells that have the stored energy resources to mount a successful metabolic response.

122

GTL

Protein Complex Analysis Project (PCAP): 3-D Reconstruction of Multi-Protein Complexes by Electron Microscopy

Bong-Gyoon Han,[1] Dieter Typke,[1] Ming Dong,[1] Pablo Arbelaez,[2] Jitendra Malik,[2] Mark D. Biggin,[1] and **Robert M. Glaeser**[1]* (rmglaeser@lbl.gov)

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

Project Goals: The broad aim of this component of PCAP is to develop high-throughput capabilities for determining the overall morphology and arrangement of subunits within large, biochemically purified multi-protein complexes of *Desulfovibrio vulgaris* Hildenborough.

Three-dimensional (3-D) reconstructions are obtained by single-particle electron microscopy (EM) at a resolution of ~2 nm for either negatively stained or unstained (cryo-EM) specimens. The goals of determining the quaternary structures of multi-protein complexes include (1) determining whether structural changes occur in some molecular machines under markedly different physiological conditions, such as those that would be encountered in the field during bioremediation, (2) providing 3-D models of their structures that can be used as templates in order to image the same multi-protein complexes within whole cells by EM tomography, and (3) using both types of information to model the biochemical networks and circuits of micro-organisms in order to better utilize them for applications in bioremediation or bioenergy.

Single-particle EM within PCAP has focused during the first two years on large soluble-protein complexes with Mr in the range 400 k to over 1000 k. These complexes have been found to differ considerably in terms of how well they hold up during EM sample preparation, and not all are stable even under the currently used conditions of cryo-EM sample preparation. Roughly half of the complexes studied have been stable enough to produce high-quality 3-D reconstructions, however, and class-average projection-images have been obtained for most of the others.

This preliminary phase of characterization has shown surprising differences in the quaternary structures of complexes isolated from *Dv*H and those that are already

* Presenting author

known for homologous proteins from other microbes. These differences occur frequently enough to make it clear that structures determined for other micro-organisms are inadequate for use as templates for modeling the biochemical networks within a given microbe of interest. By extension it is clear that the same type of EM structure determinations could be essential to characterize any changes in the multi-protein complexes that exist under different physiological conditions.

Work that is aimed towards increasing the throughput of single-particle EM currently includes the implementation of automated data collection and automated data analysis, and the engineering of new support-film technologies for EM sample preparation. The latter is driven by the need, encountered within this high-throughput project, to use technologies that do not require sample-dependent optimization and are more likely to preserve quaternary structure in a conformationally homogeneous state.

123

GTL

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

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

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

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

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

* Presenting author

tion of culturing and harvesting of wild type cells and all engineered strains are used to determine the optimal time points, best culture techniques, and best techniques for harvesting cultures using real-time analyses with synchrotron FTIR spectromicroscopy, and other methods. Finally, we are producing large quantities of cells under different conditions and harvesting techniques for optimal protein complex analyses. To insure the quality and reproducibility of all the biomass for protein complex analyses we use extreme levels of QA/QC on all biomass production. We expect to do as many as 10,000 growth curves and 300 phenotype microarrays annually and be producing biomass for 500-1000 strains per year by end of the project. Each biomass production for each strain and each environmental condition will require anywhere from 0.1 – 400 L of culture, and we expect more than 4,000 liters of culture will be prepared and harvested every year. The Microbiology Subproject is optimizing phenotyping and biomass production to enable the other Subprojects to complete the protein complex analyses at the highest throughput possible. Once the role of protein complexes has been established in the stress response pathway, we will verify the effect that the stress response has on reduction of metals and radionuclides relevant to DOE.

During the last year, the Microbiology Subproject produced biomass for multi-protein complex isolation and identification by mass spectrometry, and for imaging multi-protein complexes by electron microscopy. This year we have provided more than 2000 L of biomass consisting of more than 300 individual productions. Production volumes range from less than 1 L of DvH wt, and mutants, for imaging and development of high-throughput tagging and isolation methods, to 400 L of DvH wt for isolation of membrane protein complexes. We currently produce 100 L of DvH wt in five days, operating two 5 L fermenters in continuous flow mode in parallel. Extensive monitoring and assays are performed to ensure product quality and consistency, including continuous measurement of optical density and redox potential, and discrete sampling for AODCs, anionic composition (including organic acids), anaerobic and aerobic plating, total protein concentrations, PLFA and qPCR. The goal of incorporating different affinity or tandem affinity (TAP) protein tags into three genes to determine the best tag for use in the PCAP project has been completed. These included the Strep-tag® (IBA) for streptavidin-binding, the SPA-tag (a.k.a. CTF) that consists of a calmodulin binding motif, tobacco etch virus protease (TEV) and 3X FLAG affinity, as well as a combination of these that replaces the calmodulin binding with the Strep-tag® resulting in STF. The three genes were the dissimilatory sulfite reductase subunit C, pyruvate ferredoxin oxidoreductase subunit B and ATP synthase subunit C. Additionally, several other gene targets have been identified through close collaboration with the VIMSS/ESPP group at LBNL and are currently being tagged. To determine localization of a given gene product in the cell, we have utilized the tetracysteine, SNAP™ (Covalys) and 6XHis tags in cooperation with the EM group of the PCAP project. Currently the total number of genes tagged with CTF are 6, with STF 17, with strep 30, with tetracysteine 13 and with SNAP 11. Tagged genes were successfully generated in randomly cloned fragments of *D. vulgaris* DNA by recombineering techniques. These are being examined for introduction of the tagged genes into the *D. vulgaris* chromosome by two recombination events. This success paves the way for a HTP tagging procedure with an ordered plasmid library of *D. vulgaris* DNA fragments. To improve the plasmid insertion tagging currently used, we implemented the two-step TOPO-GATEWAY strategy (Invitrogen) for the production of a library of 145 entry clones in *E. coli*. We also constructed a library of custom destination vectors bearing the following tags: 6xHis, STF, SPA, SNAP, STF-6xHis and STF-SNAP. These destination vectors enable rapid addition of desired tags to the entry vector library. Consequently we have constructed a library of 140 STF/SPA tagged clones of which, 84 have been electroporated in *D. vulgaris* so far. Construction of the entry vector and tagged clone libraries involved development of automated software and hardware methods. On the software end we collaborated with Subgroup D (Computational Core) for the development of automated algorithms and a LIMS system for: 1) Primer identifications based on gene locations within operons for PCR amplifications in 96-well format, and 2) QA/QC for sequence data analysis and sample tracking. On the hardware end we developed and implemented methods for handling nucleic acids using a liquid handling system from Beckman Coulter and collaborated with Subgroup B (Hardware engineering) for the design of a custom electroporation device to enable rapid transformation of tagged constructs in *D. vulgaris*. All of the tagged strains constructed this year have been characterized using phenotypic microarrays (PM), and the *D. vulgaris* megaplasmid minus strain (MP(-)) being used in the electroporation studies last year was aggressively characterized for all differences including stress responses with the wildtype and was found to have some significant response differences. This enabled the group to redirect transformation studies using electroporation away from the MP(1) strain and towards the wildtype only.

* Presenting author

124

GTL

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

Swapnil Chhabra[1]* (SRChhabra@lbl.gov), Gareth Butland,[1]* Dwayne Elias,[2] **Veronica Fok**,[1] **Barbara Gold**,[1] Jian Jin,[1] Aindrila Mukhopadhyay,[1] **Ramadevi Prathapam**,[1] **Wenhong Yang**,[1] John-Marc Chandonia,[1] Judy Wall,[2] Terry Hazen,[1] and Jay Keasling[1,3]

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

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

In this poster we describe our efforts towards the development of a high throughput platform for generating a library of *D. vulgaris* mutant strains expressing tagged proteins. This work is part of the microbiology core of the Protein Complex Analysis Project (PCAP). We highlight our efforts towards automating the strain generation process using automated software and hardware tools such as LIMS for automated sequence alignments, liquid handling systems for processing nucleic acids and custom robotics for high throughput electroporations. For generation of tagged clones we have developed and tested two approaches. The first one involves the use of plasmid constructs carrying single target genes using the two-step TOPO-GATEWAY cloning approach (Invitrogen). The first step in the strategy involves generation of an entry vector carrying the gene of interest (GOI) via TOPO cloning. The second step involves transfer of the GOI from the entry vector to a suitable destination vector (carrying the tag of choice) through an in-vitro recombination reaction. This approach works best for genes located at terminal ends of operons and based on this approach we have constructed a library of 145 tagged clones in *E. coli*.

The second strategy involves the use of an ordered library of *D. vulgaris* modified using a lambda-red phage system. Library constructs are modified, in a strain of *E. coli* expressing the lambda-red recombination system, using linear PCR products specifically engineered to recombine into the 3' end of the gene of interest. These PCR products, when inserted into the gene of interest modify the coding sequence of the gene to encode a C-terminal fusion protein bearing the tag of choice. We have performed several rounds of trials to optimize the recombineering protocol and have created 15 constructs bearing individually tagged genes. This system is now being integrated into the high-thoughput tagged strain construction pipeline. The recombineering system will work in tandem with the TOPO-GATEWAY method, and will focus on genes currently not amenable to tagging via that system. Constructs generated via both strategies have been transformed into *D. vulgaris* via electroporation and are currently being tested.

125

GTL

The Center for Molecular and Cellular Systems: Biological Insights from Large Scale Protein-Protein Interaction Studies

Michelle V. Buchanan[1]* (buchananmv@ornl.gov), Dale A. Pelletier,[1] Gregory B. Hurst,[1] W. Hayes McDonald,[1] Denise D. Schmoyer,[1] Jennifer L. Morrell-Falvey,[1] Mitchel J. Doktycz,[1] Brian S. Hooker,[2] William R. Cannon,[2] H. Steven Wiley,[2] Nagiza F. Samatova,[3] Tatiana Karpinets,[1] Mudita Singhal,[2] Chiann-Tso Lin,[2] Ronald C. Taylor,[2] Don S. Daly,[2] Kevin K. Anderson,[2] and Jason E. McDermott[2]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [2]Pacific Northwest National Laboratory, Richland, Washington; and [3]North Carolina State University, Raleigh, North Carolina

Project Goals (Abstracts 126-130): The Center for Molecular and Cellular Systems (CMCS) has established a resource for high-throughput determination of protein-protein interactions (PPI) for the Genomics:GTL community.

The Center for Molecular and Cellular Systems (CMCS) has established a resource for high-throughput determination of protein-protein interactions (PPI) for

* Presenting author

the Genomics:GTL community. As part of the CMCS, an analysis "pipeline" has been established for identifying PPI among soluble proteins in *Rhodopseudomonas palustris*. The general strategy is to express an affinity-tagged protein in a bacterial culture, lyse the cells, isolate the affinity-tagged protein along with interacting proteins, and identify the affinity isolated proteins via mass spectrometry and informatics analysis. The pipeline was designed to be applicable to a wide array of gram negative bacterial species, and thus is sufficiently general to enable studies of any number of organisms that are of importance for DOE energy and environment missions. The cloning component of the pipeline is based on a flexible system (Gateway) that further expands the generality of the approach by allowing facile introduction of a wide variety of affinity (or other) tags.

The CMCS has made considerable progress toward using as affinity-tagged "baits" some 1200 *R. palustris* proteins that meet the following criteria: (1) the protein is predicted to be soluble, (2) the protein has been previously detected by mass spectrometry in proteomics studies. The results of our PPI survey in *R. palustris* are available via the Microbial Protein-Protein Interaction Database (MiPPI.ornl.gov). Statistical tools allow evaluation of the PPI based on characteristics of the data, and bioinformatics tools provide insights based on comparison of CMCS results to those from other techniques (e.g. gene expression measurements, PPI predictions) as well as PPI data from other organisms.

The results from the CMCS PPI pipeline are proving to be useful as a source of hypotheses for more detailed experiments aimed at particular pathways or systems in microbes. One example evolves around interactions observed in nitrogen fixing cells among proteins which are potentially involved in electron transfer to nitrogenase. Collaborative experiments involving the CMCS and the Harwood laboratory at the University of Washington are building on this result to explore the implications for production of hydrogen via the nitrogen fixation reaction. A further example involves study of a stress response pathway in *R. palustris* based on observed PPI involving proteins encoded by an operon that includes an ECF sigma factor, a putative response regulator, a putative histidine kinase, and an unknown protein.

Ongoing research in the CMCS is aimed at improving the throughput, applicability, and reliability of the PPI pipeline. Final validation and implementation of a robot-based protocol for affinity isolation will be completed in January 2008, removing a major bottleneck from the pipeline. Expansion of the CMCS pipeline to include membrane-associated proteins is underway. With these and other advances, the CMCS provides a unique resource for characterizing protein "machines" for the Genomics:GTL program. Details of these studies and other CMCS activities are covered in additional abstracts.

126

GTL

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

Kevin K. Anderson,[2]* William R. Cannon,[2] Don S. Daly,[2] Brian S. Hooker,[2] Jason E. McDermott,[2] Gregory B. Hurst,[1] W. Hayes McDonald,[1] Dale A. Pelletier,[1] Denise D. Schmoyer,[1] Jenny L. Morrell-Falvey,[1] Mitchel J. Doktycz,[1] Sheryl A. Martin,[1] Mudita Singhal,[2] Ronald C. Taylor,[2] H. Steven Wiley,[2] and **Michelle V. Buchanan**[1] (buchananmv@ornl.gov)

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

Project Goals: See goals for abstract 126.

The Genomics:GTL Center for Molecular and Cellular Systems (CMCS) is a DOE Center whose mission is to determine protein complexes and interaction networks for microbial systems. The CMCS is currently focusing on the completion of the characterization of soluble protein-protein interactions in *Rhodopseudomas palustris*. The CMCS approach combines expression of affinity tagged proteins, affinity purification of interacting proteins, and tandem mass spectrometric identification of these proteins. Our goal is to provide a capability for generating high quality protein-protein interaction data from a variety of energy- and environment-relevant microbial species. This poster provides a status report of the CMCS measurements of protein-protein interactions in *R. palustris*, which is of high relevance to DOE missions due to its ability to produce hydrogen, to degrade lignin monomers, and for its exceptional metabolic versatility. A critical component of the approach is our evolving data analysis pipeline.

As of early December 2007, nearly 1200 *R. palustris* genes have been cloned as Gateway entry vectors, and approximately 1060 expression clones for a dual affinity tag (6-His/V5) have been produced. Some 467 affinity-tagged bait proteins have been expressed, affinity purified,

* Presenting author

and subjected to mass spectrometry (MS) analysis to identify interacting proteins. Approximately 30% of these bait proteins are annotated as conserved hypothetical, conserved unknown, or unknown proteins.

The data pipeline for analysis of the data begins with a Laboratory Information Management System (LIMS) to capture the MS/MS data and descriptions regarding the biological and assay conditions (metadata). The LIMS maintains a detailed history for each sample by capturing processing parameters, protocols, stocks, tests and analytical results for the complete life cycle of the sample.

The resulting lists of potentially interacting prey proteins identified from MS/MS are statistically analyzed within a software environment specifically designed for working with biological networks. Bayes estimates of the confidence of the inferred associations are estimated for each bait/prey pair. For high confidence interactions, robust networks of interacting proteins are determined from patterns of interactions. The resulting protein networks are captured in a database within a publically accessible software environment (https://www.emsl.pnl.gov/SEBINI/). Using an exploratory data analysis tool that enables integration and analysis of interactions evidence obtained from multiple sources (CABIN, www.sysbio.org/capabilities/compbio/cabin.stm), the information on the nodes (proteins) and edges (interactions) can be linked to external and internal bioinformatic data. The internal bioinformatic data contains information on interologues derived from the *Bioverse* system, which provides additional information on protein interactions. The joint analysis of experimental data and multiple sources of bioinformatic data is done graphically through collective analysis of biological interaction networks (*Cabin*), a plug-in for the *Cytoscape* network visualization program.

These protein-protein interactions are disseminated through the publicly accessible Microbial Protein-Protein Interaction Database (MiPPI.ornl.gov). MiPPI is updated every 6 months (May and November). MiPPI provides tables of observed protein-protein interactions, as well as background information on CMCS measurement and analysis techniques. Various results (mass spectrometry results, corresponding metadata, and identified protein-protein interactions, including the statistical analysis scores) are also available for download in various file formats.

127

GTL

Analysis of the Dynamical Modular Structure of *Rhodopseudomonas palustris* Based on Global Analysis of Protein-Protein Interactions

William R. Cannon[2]* (william.cannon@pnl.gov), Mudita Singhal,[2] Ronald C. Taylor,[2] Don S. Daly,[2] Dale A. Pelletier,[1] Gregory B. Hurst,[1] Denise D. Schmoyer,[1] Jennifer L. Morrell-Falvey,[1] Brian S. Hooker,[2] W. Hayes McDonald,[1] **Michelle V. Buchanan**,[1] and H. Steven Wiley[2]

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

Project Goals: See goals for abstract 126.

Global determination of protein-protein interactions for *Rhodopseudomonas palustris* is the current target for the Genomics:GTL Center for Molecular and Cellular Systems (CMCS). *R. palustris* is a metabolically versatile anoxygenic phototrophic bacterium, and analyses have focused on protein interactions observed under differing conditions for nitrogen metabolism in which either NH_4^+ (fixed nitrogen) or N_2 serve as the primary source of nitrogen.

We have used the set of protein-protein interactions as the foundation for determining the dynamic modular structure of *R. palustris* regulatory networks. Global interactions determined by our affinity isolation pipeline are parsed into functional subnetworks by combining operon membership, gene regulatory information, gene expression information, phylogenetic profiling, gene neighborhood analyses and predicted interactions. Approximately 6,000 interactions between over 700 proteins were parsed in modular subnetworks and compared to the pattern of regulated gene expression observed under conditions of hydrogen utilization. We have also compared these functional modules with those inferred from protein interaction data gathered in other bacteria, such as *E. coli*. Our analysis indicates that different technologies for evaluating protein interaction networks have distinct inherent biases and that combining multiple data sources are likely to produce the most robust results. The subnetworks inferred from multiple data sources can provide novel hypotheses relating to previously unknown proteins and can serve as a foundation for further investigations—see the posters *Protein-Protein Interactions Involved in electron transfer to nitrogenase for Hydrogen Production in Rhodopseudomonas palustris* and

* Presenting author

poster *Identificataion of a Putative Stress Response Pathway and Novel Extracytoplasmic Function σ/Anti-σ Factors in the Anoxygenic Phototrophic Bacterium Rhodopseudomonas palustris by Protein-Protein Interactions* for detailed discussions of biological phenomena.

128

GTL

Characterization of a Stress Response Pathway in the Anoxygenic Phototrophic Bacterium *Rhodopseudomonas palustris*

Michael S. Allen[1]* (allenms@ornl.gov), Dale A. Pelletier,[1] Gregory B. Hurst,[1] Linda J. Foote,[1] Trish K. Lankford,[1] Catherine K. McKeown,[1] Tse-Yuan S. Lu,[1] Elizabeth T. Owens,[1] Denise D. Schmoyer,[1] Jennifer L. Morrell-Falvey,[1] W. Hayes McDonald,[1] Mitchel J. Doktycz,[1] Brian S. Hooker,[2] William R. Cannon,[2] and **Michelle V. Buchanan**[1]

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

Project Goals: See goals for abstract 126.

Rhodopseudomonas palustris is an anoxygenic phototrophic bacterium possessing high metabolic diversity. As part of the Genomics:GTL Center for Molecular and Cellular Systems (CMCS) effort, this organism has been investigated for its ability to produce nitrogenase-mediated biohydrogen and its potential for bioremediation. Analysis of cytoplasmic protein fractions by shotgun proteomics has revealed several proteins up-regulated during growth on benzoate as well as under diazotrophic conditions. Among those was the putative extracytoplasmic function (ECF) σ-factor RPA4225. Subsequent large-scale protein-protein interaction experiments also revealed an interaction between the unknown protein RPA4224 and the putative response regulator RPA4223. RPA4224 and RPA4225 form a single operon in *R. palustris*, suggesting that this unknown protein may serve as an anti-σ factor. Organization of this operon along with the preceding response regulator gene RPA4223 is conserved among several α-Proteobacteria including *Sinorhizobium meliloti*, where the components have been shown to act as mediators of the global stress response. Additionally, we have found that the genomic location of the downstream gene RPA4226, a putative histidine kinase containing a predicted transmembrane domain, is also conserved among these bacteria. This suggests a potential role in the sensing and signal transduction of the stress response. These data underscore the utility of high-throughput methodologies to interrogate complex, multi-component systems and for generating new hypotheses regarding proteins about which little or nothing is known.

129

GTL

Protein-Protein Interactions Involved in Electron Transfer to Nitrogenase for Hydrogen Production in *Rhodopseudomonas palustris*

Dale A. Pelletier[1]* (pelletierda@ornl.gov), Erin Heiniger,[3] Gregory B. Hurst,[1] Trish K. Lankford,[1] Catherine K. McKeown,[1] Tse-Yuan S. Lu,[1] Elizabeth T. Owens,[1] Denise D. Schmoyer,[1] Jennifer L. Morrell-Falvey,[1] Brian S. Hooker,[2] W. Hayes McDonald,[1] Mitchel J. Doktycz,[1] William R. Cannon,[2] Caroline S. Harwood,[3] and **Michelle V. Buchanan**[1]

[1]Oak Ridge National Laboratory, Oak Ridge Tennessee; [2]Pacific Northwest National Laboratory, Richland, Washington; and [3]University of Washington, Seattle, Washington

Project Goals: See goals for abstract 126.

The goal of the Center for Molecular and Cellular Systems (CMCS) is to identify protein-protein interaction networks that form the molecular basis of biological function in bacterial species relevant to the Genomics:GTL program. *Rhodopseudomonas palustris* is a metabolically versatile anoxygenic phototrophic bacterium that is emerging as a model system for nitrogenase-mediated biohydrogen production. This process requires the integration of several metabolic and regulatory networks, including nitrogen metabolism, photosynthesis and carbon metabolism. Although the nitrogenase enzyme has been the focus of much research, we have a poor understanding of the organization of cellular components facilitating the flow of electrons derived from carbon metabolism to nitrogenase in *R. palustris* and other diazotrophic bacteria. To better understand this and other processes, we have begun mapping the protein-protein interactions in photoheterotrophically grown *R. palustris.* Shotgun proteomics and microarray analysis have identified proteins that are upregulated in *R. palustris* cells grown in the absence of fixed nitrogen. These proteins where subsequently analyzed to identify protein-protein interactions by affinity isolation and mass spectrometry. This analysis revealed interactions among numerous proteins including FixABCX, a predicted protein complex hypothesized to have a role in transfer

* Presenting author

of electrons to nitrogenase. Subsequently we found that a *fixABCX* mutant was deficient but not completely blocked in its ability to grow under nitrogen fixing conditions. This mutant was also deficient in nitrogenase activity. Supplying fixABCX in trans restored the growth phenotype. RPA1927 and RPA1928 encode proteins of unknown function that are also highly expressed under nitrogen-fixing growth conditions. While the functions of RPA1927 and RPA1928 are unknown, the presence of a predicted ferredoxin-like iron-sulfur cluster in RPA1928 implicates this protein in electron transfer. Additionally a novel putative interaction was identified between the proteins encoded by RPA1927 and RPA1928 and the FixABCX complex implying a potential role in electron transfer. An RPA1927-RPA1928 deletion strain has been constructed and growth phenotypes are under investigation. These studies have increased our understanding of the pathways and protein-protein interactions that occur in *R. palustris* cells grown under nitrogen-fixing and hydrogen producing conditions. These results as well as the results of future interaction studies will allow for modeling and metabolic engineering of this organism for increased yields of biological hydrogen.

130

GTL

Application and Optimization of a Multi-Use Affinity Probe (MAP) Toolkit for Systems Biology

M. Uljana Mayer, Baowei Chen, Yijia Xiong, and **Thomas C. Squier*** (thomas.squier@pnl.gov)

Pacific Northwest National Laboratory, Richland, Washington

Project Goals: Newly synthesized MAPs built upon the cyanine dyes used for single molecule imaging offer the potential for multicolor measurements of protein localization and associations, and provide a path-forward for the high-throughput parallel characterization of protein-protein networks using a single tagging step and affinity technology. Because protein complexes can be released using simple reducing agents, low-affinity binding interactions can be captured, identified, and validated using the same MAPs. However, the robust utilization of MAPs requires the development of standard protocols that provide recipes and outline limitations regarding how MAPs can be used to image and purify protein complexes. We will focus on the application of existing MAPs and associated resins that we have synthesized, paying particular attention to the following deliverables. I. Demonstrate Utility of New Brighter MAPs (i.e., AsCy3) to Image Bacterial Proteins. II. Establish Ability of MAPs to Isolate Protein Complexes in Comparison with Established Tandem Affinity Purification Approaches. III. Benchmark Requirements of MAPs for Imaging and Protein Complex Measurements.

Summary: Newly synthesized multiuse affinity probes (MAPs) built upon the cyanine dyes used for single molecule imaging permit multicolor measurements of protein localization and associations, and provide a path-forward for the high-throughput parallel characterization of protein-protein networks using a single tagging step and affinity technology. Genetically encoded tags are engineered onto proteins of interest, and subsequently labeled using MAPs. The small tag size and the ability to use MAPs to image proteins under anaerobic conditions has substantial advantages relative to other technologies involving, for example, fluorescent proteins whose large size, slow folding kinetics, and requirement for molecular oxygen for chromophore biosynthesis prevent their robust application in a range of bacterial systems. Following the immobilization of MAPs on solid supports, protein complexes can be released using simple reducing agents. Low-affinity binding interactions are readily captured, identified, and validated using the same MAPs. We report the application of existing MAPs and associated resins that we have synthesized, providing examples of the following applications using the cytosolic RNA polymerase (RNAP) complex and integral membrane proteins associated with metal reduction in *Shewanella*.

I. *Demonstrated Utility of MAPs to Image Bacterial Proteins*. Live cell imaging has demonstrated the utility of using MAPs to label both cytosolic and membrane proteins in highly pigmented bacteria. The development of red/infrared MAPs provides an effective means of monitoring protein-protein interactions in highly pigmented microbes where the interference with existing chromophores interferes with the utilization of previously developed probes. Further, MAPs permit the labeling of proteins following their cellular localization, and coupled with their small size permit pulse-chase measurements of cellular trafficking.

II. *Established Ability of MAPs to Isolate Protein Complexes in Comparison with Established Tandem Affinity Purification Approaches*. Following immobilization of MAPs on solid supports, intact supramolecular protein complexes are eluted using mild reducing conditions for protein identification using mass spectrometry. In comparison with traditional tandem affinity approaches, numerous low-affinity binding interactions are retained. As the

* Presenting author

intact and functional complex is eluted, complementary structural and functional measurements are possible to assess the consequences of macromolecular organization without the need to for complex reconstitution experiments of purified proteins.

III. *Use of MAPS for the Validation of Protein-Protein Interactions and Measurements of Catalytically Important Motions for High-throughput Functional Screens*. Multicolor measurements permit the facile validation of protein-protein interactions and structural arrangements within individual protein complexes in either cellular lysates or living cells. Using defined model systems, we demonstrate the ability to identify binding interactions between proteins in complex using both fluorescence correlation spectroscopy and energy transfer measurements. Complementary measurements demonstrate the ability to assess functional protein motions that offer a means to assess changes in protein function in living cells in response to changes in environmental conditions.

131

GTL

Development of Highly Efficient Bacterial Hosts for High Throughput Recombinant Membrane Protein Production

Hiep-Hoa T. Nguyen* (hiephoa@its.caltech.edu), Sanjay Jayachandran, and Randall M. Story

TransMembrane Biosciences, Pasadena, California

Project Goals: The objective of this project is to develop superior host strains for efficient and high yield production of correctly folded and functional recombinant membrane proteins.

Membrane proteins and enzymes (for instance channels, receptors, and transporters) are involved in critical cellular processes, but our understanding of these important biological molecules at molecular levels falls behind those of soluble proteins. For example, while thousands of X-ray crystal structures of soluble proteins are known, only ~80 structures of unique membrane proteins are currently available. This disparity is increasing with the surge in data generated by various genome and structural genomics projects. The most significant problem precluding any structural characterization of membrane proteins is their low levels of biosynthesis. As a result, obtaining significant quantities (even at milligrams scale) of purified membrane proteins for biochemical and biophysical studies has been a major obstacle. An effective membrane protein overexpression system would be indispensable and allow us to tackle this intractable but very significant problem in membrane protein biochemistry. All of the commercially available protein expression vehicles yield very poor results for membrane proteins even though the most powerful systems can generate *gram* quantities of recombinant soluble proteins even with small scale fermentation. Considering the ineffectiveness of current expression systems in overexpressing membrane proteins, it is clear that economical and effective expression systems for membrane proteins are needed. The objective of this proposal is to develop high yield membrane protein expression systems. Specifically, we seek to obtain superior bacterial hosts for membrane protein production. The goal of this research is to improve further the yield of recombinant membrane proteins produced by our bacterial hosts to the extent that significant quantities can be obtained with small scale cultures. We have pursued this objective through a combination of rational genetic engineering and directed evolution/screening process to obtain superior membrane protein production hosts using *Escherichia coli*, a familiar, robust, and highly amendable microorganism. We have been working in creating strains that overexpress proteins involving in membrane protein biosynthesis using background mutants obtained through screening efforts. The background mutants were selected through screening assays that demonstrate the mutants' capability to tolerate high expression levels of recombinant membrane proteins. Evaluation of the improved strains obtained through genetic engineering by test-expression of a number of model recombinant membrane proteins is in the way. As a demonstration for the power of our expression technologies, recently, we produced >100 mg of purified wild type and Se-Met labeled Rh protein (a channel with 11 transmembrane helices) within 2 months and solved the X-ray crystal structure of this channel in 5 months. The 1.8 A resolution X-ray crystal structure of this channel and related results were recently published in *PNAS*. The speed and efficiency of this effort is a validation for our approach in membrane protein production.

* Presenting author

Validation of Genome Sequence Annotation

132

GTL

A High Thoughput Proteomic and Protein Expression Strategy for Annotation of Fungal Glycosyl Hydrolases

Scott E. Baker[1]* (scott.baker@pnl.gov), Jon K. Magnuson,[1] Ellen A. Panisko,[1] Adrian Tsang,[2] and Frank Collart[3]

[1]Fungal Biotechnology Team, Energy and Environment Directorate, Pacific Northwest National Laboratory, Richland, Washington; [2]Fungal Genomics, Concordia University, Montreal, Canada; and [3]Biosciences Division, Argonne National Laboratory, Lemont, Illinois

Project Goals: The objective of this project is to use high throughput proteomics, protein expression and enzyme assays to rapidly generate functional data and annotation for secreted proteins, specifically glycosyl hydrolases.

There has been an exponential increase in the number of microbial genomes sequenced but not the number of methods for functional annotation. These still rely primarily on sequence similarity whether performed by algorithms or by manual curators. This is very effective with enzymes that have been well characterized and exhibit a high degree of sequence identity across phyla, such as central metabolic enzymes. However, when insufficient experimental information on substrate specificity is available, only general functions can be assigned to the genes, or worse, improper annotations can be made and quickly propagated through the databases. This is especially true for glycosyl hydrolases as groups of families based on sequence similarities often have a variety of substrate specificities. The crucial role of fungi in environmental carbon and nitrogen cycling stems from their absorptive nutritional strategy. The glycosyl hydrolases that fungi secrete are integral to this strategy and in addition, comprise a large fraction of fungal genes. Filamentous fungal genomes contain between 100 and 220 glycosyl hydrolase genes for the breakdown and utilization of plant, fungal and prokaryotic cell walls. Glycosyl hydrolases (GH) are critical for the hydrolysis of plant derived biomass for subsequent fermentation to liquid fuels such as ethanol as well as products such as organic acids. There are over 150 recognized GH activities in the enzyme classification system (EC 3.2.1._), and this category is growing. The CAZy database (http://afmb.cnrs-mrs.fr/CAZY/) is an excellent source of information about glycosyl hydrolases (and other carbohydrate active enzymes). The GHs are currently divided into over 100 families based on sequence similarity and these families are extremely helpful for annotation of genomes. However, even within families, multiple activities are often found, such that primary sequence homology alone is insufficient to definitively assign a function. Therefore, we are developing a high throughput protein-centric annotation pipeline for fungal GHs (Figure) consisting of mass spectrometry based proteomic identification of secreted proteins from fungi grown on a variety of biomass substrates, high throughput protein expression of fungal GHs prioritized by the proteomic analyses, and substrate development for multiplexed assay of native and recombinant GHs. A variety of substrates and assay methods will be developed for high-throughput functional classification and characterization of fungal secreted enzymes. The focus will be on known and potentially novel glycosyl hydrolase families.

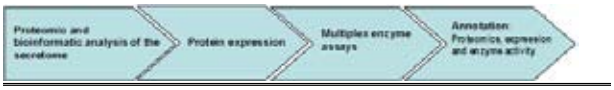


Figure. Pipeline for high throughput functional annotation of fungal glycosyl hydrolases.

133

GTL

Assignment of Enzymatic Function for Core Metabolic Enzymes

Vincent Lu,[1]* Gopi Podila,[2] Michael Proudfoot,[3] Alexander Yakunin,[3] and **Frank Collart**[1] (fcollart@anl.gov)

[1]Biosciences Division, Argonne National Laboratory, Lemont, Illinois; [2]Department of Biological Sciences, University of Alabama, Huntsville, Alabama; and [3]Banting and Best Department of Medical Research, University of Toronto, Toronto, Canada

Project Goals: 1. Functionally map the set of conserved hypothetical proteins from *Shewanella oneidensis* which contains ~800 members (TIGR annotation). The rationale for screening the set of conserved hypothetical from a single genome is to benchmark the utility of the enzymatic screening approach for the improvement

* Presenting author

of functional assignment for a large set of proteins of unknown function. 2. The second aim will be to apply a directed screening strategy to uncharacterized proteins of the haloacid dehalogenase (HAD)-like hydrolase superfamily, which will be tested for the presence of dehalogenase, phosphonatase, phosphatase, or glucomutase activities for HAD-like hydrolases. This aim will use several strains of *Shewanella* and the symbiotic fungus *Laccaria bicolor*. This objective will provide a foundation to assess the capabilities for specific functional assignments for a substantial number of unknown prokaryotic and eukaryotic proteins.

With over 600 genomes with complete sequences currently available in public databases and thousands of genome sequence projects in progress, there's a pressing need to effectively annotate genomic sequences quickly and accurately for functional activity. The main objective of this proposal is to experimentally annotate (assign a biochemical function) a large group of conserved hypothetical proteins using high throughput protein production and enzymatic screening methods. This approach for experimental annotation will be applied to hypothetical proteins from a prokaryote and a eukaryote of programmatic interest. In the first stage of the project we will functionally map the set of conserved hypothetical proteins from *Shewanella oneidensis* which contains ~800 members (Figure). The rationale for screening the set of conserved hypothetical from a single genome is to benchmark the utility of the enzymatic screening approach for the improvement of functional assignment for a large set of proteins of unknown function. A second component of the project will be to apply a directed screening strategy to uncharacterized proteins of the haloacid dehalogenase (HAD)-like hydrolase superfamily, which will be tested for the presence of dehalogenase, phosphonatase, phosphatase, or glucomutase activities for HAD-like hydrolases. This aim will use several strains of *Shewanella* and the symbiotic fungus *Laccaria bicolor*. Targets from *L. bicolor* will be amplified from cDNA clones, clone libraries or generated using a PCR-based gene synthesis approach. This objective will provide a foundation to assess the capabilities for specific functional assignments for a substantial number of unknown prokaryotic and eukaryotic proteins.

For protein production, we will use the efficiency of automated strategy to implement a parallel pipeline consisting of an *E. coli* and yeast expression systems. The screening strategy uses a tiered approach where targets are categorized using a series of general screens and then rescreened for specific functional assignments using a directed series of natural substrates (Figure). The general screening assays have relaxed substrate specificity and are designed to identify the subclass or sub-subclasses of enzymes (phosphatase, phosphodiesterase/nuclease, protease, esterase, dehydrogenase, and oxidase) to which the unknown protein belongs. Further biochemical characterization of proteins can be facilitated by the application of secondary screens with natural substrates (substrate profiling).

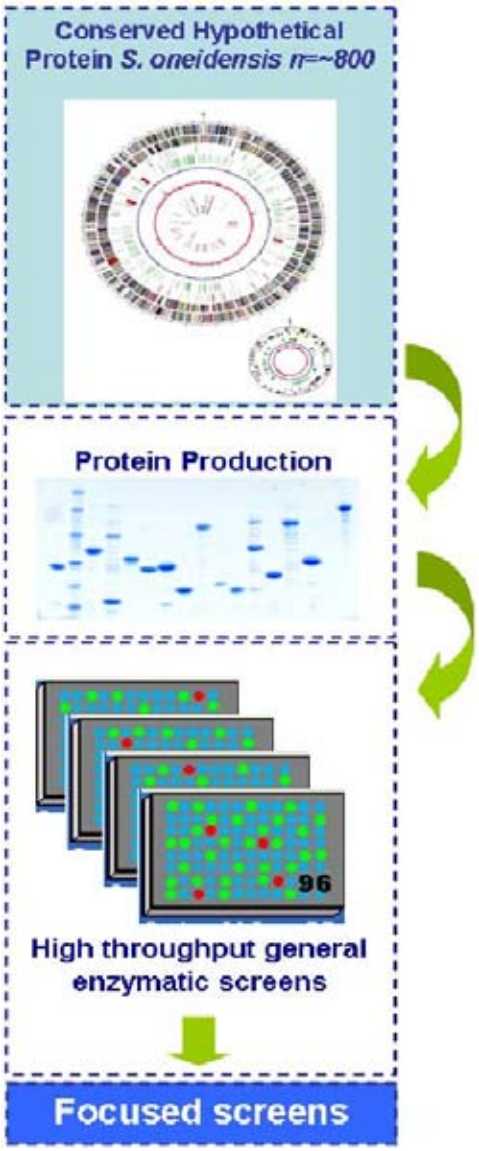


Figure. Illustration of the protein production and enzymatic screening process.

* Presenting author

134

GTL

Characterization of Sensor Proteins and Domains

Sarah Giuliani,[1]* Lori Field,[1] F. William Studier,[2] Lisa M. Miller,[3] and **Frank Collart**[1] (fcollart@anl.gov)

[1]Biosciences Division, Argonne National Laboratory, Lemont, Illinois; [2]Biology Division, Brookhaven National Laboratory, Upton, New York; and [3]National Synchrotron Light Source, Brookhaven National Laboratory, Upton, New York

Project Goals: 1. Specific functional assignments for several classes of sensor proteins. 2. Functional assignments for a set of homologs that can be used to define sequence motifs that can improve annotation assignments based on sequence alignments. 3. An expression clone library with extensive characterization data. 4. A novel set of versatile vectors customized for expression of sensor proteins.

All cells contain proteins that sense the environment and mediate transport and signaling events that lead to changes in metabolism and/or initiate changes in gene expression at the level of transcription. Mapping of ligands with these binding/sensor proteins is critical to our understanding of cell biochemistry and is essential for modeling cellular processes and the rational design of engineered organisms. The goal of this project is to evaluate a series of experimental techniques to screen for potential ligands that bind to "sensor type" proteins. The experimental approach is based on the observation that ligand binding for many proteins can alter stability of the protein. A fluorescence-based thermal shift assay was used for the identification of bound ligands and assignment of function. This is a target independent assay that uses a florescent dye to monitor protein unfolding and has been widely used for the assessment of ligand binding. This assay uses a commercially available real-time PCR instrument where thermal melting curves of the protein/ligand combinations can be screened in a 96-well plate format. To illustrate the suitability of this approach for ligand binding, we used Carbonic Anhydrase I [CAI, (EC 4.2.1.1)] as a positive control test protein. Our analysis indicated that the native form of this protein displays a concentration-dependent thermal shift (Fig. 1a). Addition of the tight binding inhibitor trifluoromethanesulfonamide (TFMSA) results in an increase in the Tm indicating enhanced protein stability (Fig. 1b). This approach will be used to improve gene/protein functional assignments of sensor type proteins by developing high throughput methods to match ligands with their binding proteins. A library of preliminary candidate binding ligands was generated using structural models in the Protein Data Bank (PDB) to identify sensor type proteins containing bound ligands. Our results show there are many structures with unique combinations of sensor type proteins and ligands. The bound ligands can be grouped into several categories such as amino acids, metals, small ions, sugars, and vitamins. Protein sequences derived from the PDB set will be used to identify potential homologs in a set of reagent genomes as candidates for functional screening. Targets for screening will be produced at Argonne using an establish pipeline for protein production and will be extensively characterized as assurance of protein or domain structural integrity which is necessary prior to ligand screening. A parallel effort at Brookhaven National Lab led by Dr. Studier will test potential strategies for increasing the efficiency of producing soluble, functional proteins, which could then be integrated into the Argonne protein production pipeline. This project will provide specific functional assignments for an important class of regulatory molecules and generate an expression-clone resource for future structural and functional studies.

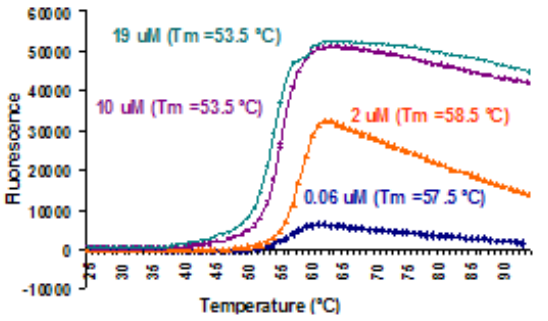


Fig. 1a. Protein concentration dependence of thermal shift assay with CAI and 5X SYPRO orange dye.

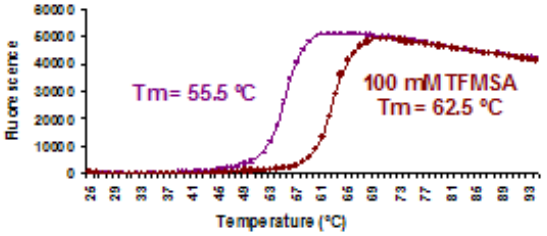


Fig. 1b. Thermal shift assay using 10 uM CAI and 5X SYPRO orange dye with and without ligand (100mM TFMSA)

* Presenting author

135

GTL

Protein Annotation from Interaction Networks using Zorch and Bayesian Functional Linkages

Richard Llewellyn[1]* and **David Eisenberg**[2] (david@mbi.ucla.edu)

[1]UCLA-DOE Institute for Genomics and Proteomics and [2]Department of Chemistry and Biochemistry, University of California, Los Angeles, California

Project Goals: Our goal is to provide a rigorous platform, known as Bayesian functional linkages, to integrate various types of functional inferences into a single description of the biological role of an uncharacterized target protein. The resulting annotation is a product of integrating the biological processes of the characterized functionally-linked proteins with any existing knowledge about the cellular location or molecular function of the target protein. We demonstrate its performance in yeast with links defined by zorch, a measure of connectivity of proteins in an interaction network (DIP). A central part of our goal is to identify and control sources of error in computational analysis so that as much information as possible can be used to infer the function of unknown proteins without diluting the accuracy of results. Bayesian functional linkages is a generalized method that we can extend to annotating proteins from uncharacterized prokaryotic genomes by combining inferences from homology with genome context and phylogenetic profiles.

Functional linkages describe relationships between proteins that work together to perform a biological task. Here we develop a general framework known as Bayesian functional linkages to annotate target proteins (Figure). We demonstrate its performance with links defined by zorch, a measure of connectivity of proteins in an interaction network. We started with the protein-protein interaction networks archived in the Database of Interacting Proteins (DIP).

Our goal is to provide a rigorous platform to integrate various types of functional inferences into a single description of the biological role of a protein. The resulting annotation is a product of integrating the biological processes of the linked predictor proteins with any existing knowledge about the cellular location or molecular function of the target protein. In Bayesian terminology, the data are the given annotations of the predictor proteins, the type of evidence supporting these annotations, and the strength of each functional linkage to the target. The hypotheses are the possible roles of the target protein as described by Gene Ontology biological process annotations. We address several of the common challenges of annotation by functional linkage: the contribution of each linked predictor protein is modulated by both the strength of its linkage to the target and the confidence that its characterized functions are correct according to their supporting evidence, in a manner that accounts for predictors with multiple annotations. The Bayesian likelihood focuses on the predictor annotations that most strongly support each hypothesis and quantifies the observation that different types of functional linkages are more likely to link proteins with particular types of functions. The contribution of any prior knowledge of the target's cellular location or molecular function is balanced against the number and quality of available functional linkages, with the final prediction given either as a posterior distribution over annotations that retains the influence of competing hypotheses, or as a single Bayes classifier of the most probable, general description of the biological process of the target.

We show that functional linkages quantified by zorch are good predictors of the biological process of proteins in *Saccharomyces cerevisiae*, and that the lack of high quality links can often be mitigated by the use of many weak ones, e.g. those inferred only from indirect or high-throughput protein-protein interactions, so that functional annotation can be extended to uncharacterized yeast proteins that lack reliably determined interacting partners.

A central part of our goal is to identify and control sources of error in computational analysis so that as much information as possible can be used to infer the function of unknown proteins without diluting the accuracy of results. We estimated the error in the existing GO annotations of predictor proteins when supported by different evidence types, such as Inferred from Expression Profile or Reviewed Computational Analysis, by comparing the accuracy rate of predicting known proteins annotated via gold standard evidence (i.e. Inferred from Direct Assay) from a linked protein with varying types of evidence. We also develop a novel method of treating GO annotations as hypotheses that explicitly addresses the incompleteness of any current ontology.

Bayesian functional linkages is a generalized method that we can extend to annotating proteins from uncharacterized prokaryotic genomes by combining inferences from homology with genome context and phylogenetic profiles. We expect that our focus on combining different types of inference while controlling sources of error will provide more accurate and higher resolution annotations

* Presenting author

than methods that treat functional linkages separately from homology.

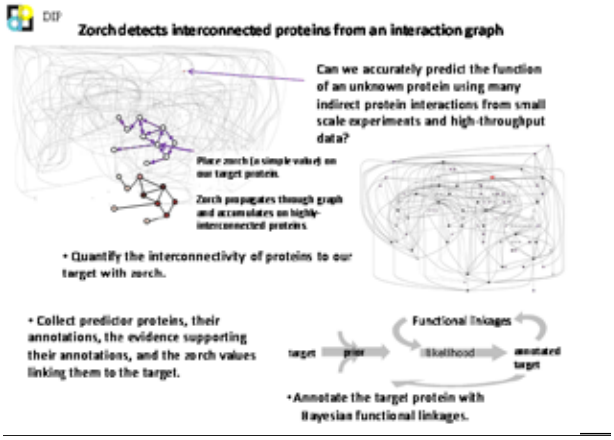


Figure. Overview of protein annotation with zorch and Bayesian functional linkages. Proteins interconnected to the unknown target are identified with zorch and integrated with any existing prior knowledge to predict the biological process.

We thank DOE BER for support.

136

GTL

Annotation of Novel Enzymatic Functions in Methanogens

Ethel Apolinario,[1] Zvi Kelman,[2] Jing Li,[3] Basil J. Nikolau,[3] Kevin Sowers,[1] and **John Orban**[2]* (orban@umbi.umd.edu)

[1]Center of Marine Biotechnology, University of Maryland Biotechnology Institute, Baltimore, Maryland; [2]Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, Maryland; and [3]W. M. Keck Metabolomics Research Laboratory, Iowa State University, Ames, Iowa

Project Goals: An integrated high-throughput approach is being developed to functionally annotate a large group of poorly understood genes in the methanogenic archaeon, *Methanosarcina acetivorans*. The focus is on genes predicted to encode enzymes, the substrate(s) and products of which are unknown. Approximately 2226 of the 4524 genes in *M. acetivorans* fall into this category and include genes possibly involved in processes such as methanogenesis, nitrogen fixation, and carbon assimilation. The biochemical functions of these putative enzymes will be accurately annotated using a combination of gene knockouts, high throughput metabolomic analysis with mass spectrometry (MS), automated screening of implicated metabolites with nuclear magnetic resonance spectroscopy (NMR), and biochemical assays.

Targets for study will include uncharacterized genes that have been associated with metabolic pathways by transcript expression using genomic microarrays. Also gene products with sequence or structural homology to proteins with a putative enzymatic function will be selected as targets. Gene disruption will be conducted by complementing proline auxotrophy in a *proC*- strain of *M. acetivorans*, which allows later complementation with a plasmid containing the wild type gene and puromycin resistance (pac) cassette. Genes will be disrupted by insertion of *proC* into the target open reading frame (ORF) followed by homologous recombination into a proline auxotroph of *M. acetivorans*. In cases where multiple gene disruptions are necessary, a directed markerless genetic disruption system will be used.

Using MS-based analyses, metabolic profiles will be compared between isogenic strains that carry either wild type or gene-knockout alleles at the locus of interest. Changes in metabolite pools in knockouts will provide clues about the specific pathway or set of pathways relevant to the function of the gene target. The extent to which MS data alone will infer a function will depend on whether there are any other intersecting pathways that may shunt elevated metabolites off in other directions.

To further increase the probability of obtaining useful functional annotations, we propose to screen potential substrates and products, or their structural analogs, from affected pathways for interaction with the putative enzyme in question using NMR methods that are not limited by molecular weight considerations. Metabolites and analogs will be obtained from commercial sources and proteins needed for screening will be expressed and purified in-house.

Compounds that interact with the protein of interest will be tested in a defined biochemical enzyme assay to validate the functional annotation. For example, if a putative enzyme is thought to be a methyltransferase then the compounds/metabolites that interact with it based on combined MS/NMR data will be used in an assay to determine whether a methyl group can be transferred to them. A range of assays will be utilized and MS and NMR will also be employed to monitor turnover reactions.

Preliminary results illustrating our general approach will be presented.

* Presenting author

137

GTL

Genemap-MS: High Throughput Mass Spectrometry Methods for Functional Genomics

Trent R. Northen[1,2]* (trnorthen@lbl.gov), Linh Hoang,[2] Steven M. Yannone,[1] Jason Raymond,[4] Jill Fuss,[1] Jinq-Chyi Lee,[3] Der-Ren Hwang,[3] Chi-Huey Wong,[3] John Tainer,[1,2] and **Gary Siuzdak**[1,2]

[1]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; [2]Department of Molecular Biology and Scripps Center for Mass Spectrometry and [3]Department of Chemistry and the Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, California; and [4]School of Natural Sciences, University of California, Merced, California

Project Goals: The utility of genetic information being derived from sequencing efforts is diminished by the incomplete and sometimes incorrect annotations associated with "completed" genomes. Homology-based protein function predictions are limited by evolutionary processes that result in conserved domains and sequence being shared by enzymes of widely diverse functions. Therefore, additional experimental datasets directed at validating and improving genome annotations are required. Here an integrated approach is used to develop universally applicable high-throughput (HT) methods for validating genome annotation using mass spectrometry (MS) based proteomics, metabolomics, and our developing technologies for detecting biochemical activities on arrayed metabolite substrates. This leverages expertise with developing and applying MS technologies to generate datasets directly applicable to validation. These general and multifaceted approaches focus on: 1) protein expression 2) metabolic pathways and 3) biochemical activities. The datasets from these analyses are integrated into computational metabolic networks to provide physical and functional validation of the many hypothetical and predicted proteins and activities in current genome annotations. We provide a balance of mature and robust MS technologies with new surface based MS technologies and the expansion of our METLIN metabolite database with the ultimate goal of addressing specific DOE needs for exploiting microbial processes for bioenergy production.

Mass spectrometry's ability to efficiently generate intact biomolecular ions in the gas phase has led to its widespread applications including metabolomics, proteomics, and biological imaging. Matrix-Assisted Laser Desorption/Ionization (MALDI) and Electrospray Ionization (ESI) have been at the forefront of these developments. We recently introduced Nanostructure-Initiator Mass Spectrometry (NIMS), a sensitive new tool for spatially defined mass analysis which complements existing methods by enabling the analysis of metabolites from tissues, cells, microarrays, and etc. with high sensitivity. NIMS utilizes 'initiator' molecules trapped in nanostructured surfaces or 'clathrates' to release/ionize intact molecules adsorbed from the surface. This technology has recently been extended for the direct screening of cell lysates and environmental samples for enzymatic activities at high temperatures and low pH values. Using this approach a new thermophilic galactosidase was identified from a Yellowstone hot springs microbial community. In addition the optimal pH, temperature, and enzyme inhibition were screened *in situ.* This general approach provides an efficient method for screening environmental sample prior to sequencing and cloning efforts and without obtaining pure cultures.

138

GTL

The Application of Phage Display to Advanced Genome Annotation: *C. Thermocellum* as an Example

Andrew Bradbury* (amb@lanl.gov)

Los Alamos National Laboratory, Los Alamos, New Mexico

Folding reporters are proteins with easily identifiable phenotypes, such as antibiotic resistance or fluorescence, whose folding and function is compromised when fused to poorly folding proteins or random open reading frames. We have found that when DNA fragments are fused to a ß lactamase folding reporter, selection for fragments of real genes, as opposed to random ORFs, tends to occur. *We hypothesize that folding reporters can be used on a genomic scale to select collections of correctly folded protein domains from the coding portion of the DNA of any organism. This technology will be applicable to any intronless genome, or collection of open reading frames, without extensive analysis or primer synthesis. This can be considered to be the protein equivalent of shotgun sequencing, and could be termed the "domainome", to extend an overused cliché.* It is expected that the protein fragments obtained by this approach will be well expressed and soluble, making them suitable for structural studies, antibody generation, protein/substrate binding analyses, domain shuffling for enzyme evolution and protein chips.

* Presenting author

Of the many functions undertaken by modular domains, molecular recognition is the most easily assessed, and the most straightforward to translate into gene annotation. By cloning the "domainome" directly in a phage display context, it will be possible to select gene fragments encoding domains with specific binding properties (e.g. to other proteins, domains, metabolites, enzyme substrates), *providing essential experimental information for gene annotation.* We expect this concept can be extended to activity based probes (ABPs)[1], to identify domains with catalytic activities. Once a domain fragment library has been created, it is a renewable resource, easily retested against new potential binding partners or ABPs.

This hypothesis will be tested with DNA from the genome of *C. thermocellum* and specifically applied to the identification of cellulose binding domains (CBDs). This genome is such a rich source of cellulase genes that it provides an excellent model system for the proposal described here, with the possibility of experimentally identifying novel cellulases or other enzymatic activities linked to CBDs, and subsequently extending it to the analysis of cohesions and dockerins.

In this poster the concept, and the progress we have made with the display of cellulose binding domains derived from *C. thermocellum* will be described.

139

GTL

Phylogenomics-Guided Validation of Function for Conserved Unknown Genes

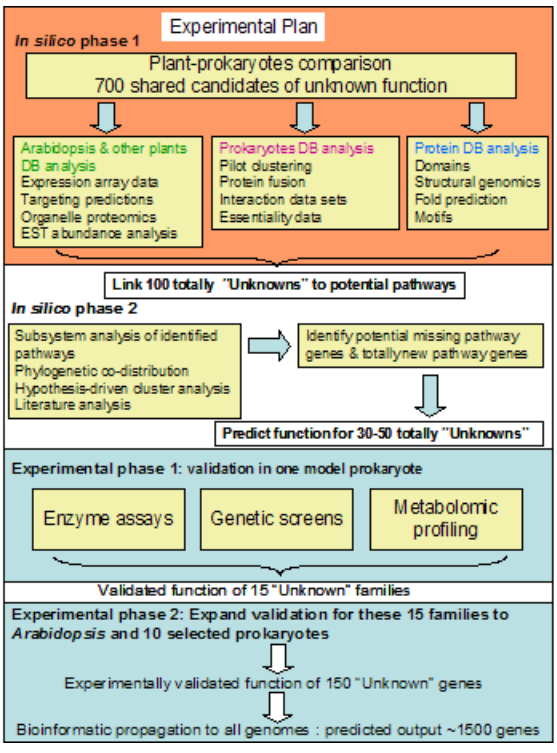
Valérie de Crécy-Lagard[1]* (vcrecy@ufl.edu), Basma El Yacoubi,[1] Crysten Haas,[1] Valeria Naponelli,[2] Alexandre Noiriel,[2] Jeffrey C. Waller,[2] and Andrew D. Hanson[2]

[1]Department of Microbiology and Cell Science and [2]Department of Horticultural Sciences, University of Florida, Gainesville, Florida

Project Goals: Identifying the function of every gene in all sequenced organisms is a major challenge of the post-genomic era. Our objective is to use an integrative approach to predict and experimentally verify the in-vivo function of genes that lack homologs of known function ('unknown' gene families) and that are highly conserved among prokaryotes and plants.

The approach has four phases. A phylogenomic analysis comprising two *in silico* phases will lead to prediction of function for 30-50 unknown gene families. Then come two experimental phases: validation of the prediction in one prokaryote and *Arabidopsis*, then extension of the validations to ten other organisms using phenotypic and enzyme assays developed in the previous phase. We expect the final outcome to be experimentally validated functions for 15 families of unknowns, which translates into ~150 individual genes. These functions can then be propagated with confidence to all genomes, leading to the functional annotation of an estimated 1500 genes.

The phylogenomic analysis is ongoing, but in pilot work we have predicted the general function of around ten families. Further bioinformatic analysis combined with experimental validations has led to more precise functional prediction for four of these families. These will be presented in more detail and consist of: 1) The universal YrdC/Sua5 family (COG009) that is involved in the modification of tRNA; 2) The CobW family (COG0523) that is a metal chaperone that might have a important role in zinc homeostasis. 3) YgfZ, a folate binding protein that could be involved in repair of iron sulfur proteins; 4) The pterin carbinolamine dehydratase family (COG2154), which occurs in many organisms that lack the pterin recycling pathway requiring the dehydratase and that could have another function in molybdenum cofactor maintenance.



* Presenting author

140

GTL

Prodigal: A New Prokaryotic Gene Identification Program with Enhanced Translation Initiation Site (TIS) Prediction

Doug Hyatt, **Loren Hauser*** (hauserlj@ornl.gov), Frank Larimer, and Miriam Land

Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee

Project Goals: Enhanced quality of gene prediction and annotation in JGI microbial organisms.

High-quality annotation of microbial genomes remains an ongoing challenge for the Joint Genome Institute. For the past several years, Oak Ridge National Laboratory (ORNL) has aided in gene prediction and functional analysis of numerous microbial organisms. As a result of the comprehensive review and curation of a large number of genomes ranging from low-GC to high-GC and from bacteria to archaea, numerous areas of improvement have been found. Predicting the correct number of genes, reducing the number of false positives, correctly locating short and laterally transferred genes, performing robustly in high-GC-content genomes, and accurately finding the translation initiation site (TIS) of genes continue to be challenges that will enhance the quality of the final JGI annotations submitted to Genbank and placed in IMG.

Prodigal (Prokaryotic Dynamic Programming Genefinding Algorithm) was developed at ORNL to address many of the "real-world" challenges discovered through many hours of manual curation of microbial genomes. In particular, the previous pipeline (based on the genefinders Critica and Glimmer) often lengthened genes in high-GC content genomes and incorrectly omitted genes that would overlap the erroneously long genes. We decided to address this issue with a new gene identification algorithm that would perform robustly in high-GC content genomes. Prodigal's self-training methodology is based on a detailed analysis of the GC-frame-plot of the organism in question. The training process consists of determining the statistical significance of G and C in different frame positions and performing a dynamic programming algorithm using this information to construct an initial training set of genes. This is a novel approach compared to other programs, which construct their training sets based merely on all open reading frames (ORFs) above a particular length. Our implementation of a coding scoring function based on this training set was found to perform well in both low-GC and high-GC genomes.

The other improvement to the annotation pipeline is improved start site prediction. Prodigal contains a novel method for examining the upstream regions for ribosomal binding site (Shine-Dalgarno) motifs. The statistical significance of various motifs relative to the background is determined automatically by an iterative algorithm which learns the organism's preference for various RBS motifs. Results for this enhanced start site prediction are presented, as well as overall results for locating the 3' end of genes. In addition, future improvements to the algorithm are discussed, such as validation through proteomics data and improvement of start site prediction via signal peptide information.

* Presenting author

Computing Resources and Databases

141

GTL

Further Refinement and Deployment of the SOSCC Algorithm as a Web Service for Automated Classification and Identification of *Bacteria* and *Archaea*

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

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

Previously, we had demonstrated that techniques such as principal components analysis (PCA), could be useful in unraveling discontinuities between classical taxonomic views of *Bacteria* and *Archaea* and phylogenetic models based on the 16S rRNA gene or other universally applicable molecular signals. PCA is a highly robust and efficient unsupervised method of data analysis that could be applied to very large sequence datasets (>10,000) and readily allow visualization of phylogenetic data in a manner that has decided advantages over classical treeing methods. Our early studies revealed that the dimensionality of such data could be reduced to two to three dimensions without a significant loss of information, allowing us to gain insight into the phylogenetic topology defined by the 16S rRNA gene. We also discovered that there were numerous anomalies between the classical taxonomic view of *Bacteria* and *Archaea* and the phylogenetic view that were largely attributable to unresolved synonymies which appeared as outliers in 2D and 3D projections. While useful, PCA was found to have inherent limitations; notably a natural weighting that was attributable to larger taxonomic groups, a variable degree of distortion and rotation that is attributable to the manner in which the reduced dimensions are calculated, and an inability to determine the correct placement of outliers, especially when those outliers were members of minor taxonomic groups.

Subsequently, we discovered that evolutionary distance matrices could be readily viewed as heatmaps within the S+ and R statistical computing environments. This approach had decided advantages over PCA as it allowed us to view the complete data matrices in a distortion-free manner, to visualize different taxonomic arrangements of the data, to pinpoint and correct nomenclatural errors, and to determine what actions were needed to bring the classical taxonomic and phylogenetic views into closer agreement. This phase of the project led to the development of a self-organizing self-correcting classification (SOSCC) algorithm that could pinpoint such anomalies in a semi-automated manner and then optimize the underlying matrix to resolve these anomalies. A prototype was developed in S+ based on the SOSCC algorithm and applied to solve a number of taxonomic problems that have accumulated in the burgeoning 16S rRNA data set, as well as to provide insight into the requirements for deploying this technique as a web service through the RDP.

As the project evolved, it became obvious that neither S+ nor R were sufficiently stable environments for building a client application. To that end, we have re-implemented the SOSCC algorithm in java as an xfire service, optimized the algorithm to provide a more satisfactory user experience (e.g. 30 seconds to produce a maximally smoothed matrix of 1000 sequences), and gained insight into a poorly understood limitation of previous versions of the SOSCC, in which correct placement of some sequences could not be achieved when the algorithm was run in a fully unsupervised, automated version. Work continues on re-implementing the presumptive identification and automatic renaming features of the original algorithm, which will be used to update the next release of the Taxonomic Outline of Bacteria and Archaea and will link the heatmap visualizations to NamesforLife information objects. Deployment of a beta-version SOSCC as an RDP service is planned for the first quarter of 2008. Once deployed, this service will provide valuable information on disagreements between classifications and phylogenies of the prokaryotes and how these problems might be resolved. Furthermore, this experience will provide insights as to how the methodology might be applied to prokaryotes and eukaryotes using other molecular sequences (including complete genomes). The SOSCC may also prove useful in providing insight into expression profiles by pinpointing similarities or discontinuities in microarray data, displayed as optimally smoothed matrices.

* Presenting author

142

GTL

NamesforLife Resolution Services for the Life Sciences

George M. Garrity[1,2]* (garrity@msu.edu), Catherine M. Lyons,[2] and James R. Cole[1,2]

[1]NamesforLife, LLC, East Lansing, Michigan and [2]Michigan State University, East Lansing, Michigan

Project Goals: NamesforLife (N4L) is an information technology that persistently resolves ambiguity in terminology through the use of a proprietary data architecture that is coupled with persistent identifiers (in the current implementation Digital Object Identifiers are used) and expertly managed terminologies. N4L technology provides transparent links to the occurrence of a technical term or biological name in third party databases or electronic content to managed information about the origins of the term, formal definition, current usage, and related goods and services.

Within the Genomes-to-Life Roadmap, the DOE states that a significant barrier to effective communication in the life sciences is a lack of standardized semantics that accurately describe data objects and persistently express knowledge change over time. As research methods and biological concepts evolve, certainty about correct interpretation of prior data and published results decreases because both become overloaded with synonymous and polysemous terms. Ambiguity in rapidly evolving terminology is a common and chronic problem in science and technology.

NamesforLife (N4L) is a novel technology designed to solve this problem. The core of the technology is an ontology, an XML schema, and an expertly managed vocabulary coupled with Digital Object Identifiers (DOIs) to form a transparent semantic resolution service that disambiguates terminologies, makes them actionable, and presents them to end-users in the correct temporal context. In the first instance, N4L technology has been applied to biological nomenclature, specifically the validly published names of Bacteria and Archaea. These names play a significant role in science, medicine, and government, carry specific meanings to end-users in each of those communities, and can trigger responses that may or may not be appropriate. Biological names also serve as key terms used to index and access information in databases and the scientific, technical, medical, and regulatory literature. Clear understanding of the correct meaning of a biological name, in the appropriate context, is essential. This is a nontrivial task, and the number of individuals with expertise in biological nomenclature is limited. This knowledge can, however, be accurately modeled and delivered through a networked semantic resolution service. Such a service could provide end-users of biological nomenclatures or other dynamic terminologies with the appropriate information, in the correct context, on demand. The same service could also be used by database owners, publishers, or other information providers to semantically enable their offerings, making them discoverable, even when the definition of a name or term has changed.

As proof of principle, a working model of the N4L technology has been built. It allowed us to validate our concepts and gain new insights into previously unaddressed complexities of dynamic vocabularies. The working model also allowed us to introduce the technology to businesses that rely on the proper use of biological names in their product offerings, including scientific publishers, instrument vendors, and other suppliers of information and biological materials. This provided us the opportunity to explore how N4L technology could be applied in various commercial settings to fulfill unmet business needs of vendors and their customers and to do so in a self-supporting manner. The latter goal is achievable because the N4L data architecture is generalizable. The problem that biologists face with terminology, whether it relates to an organism, a gene, or a gene product, is not unique. Analogous problems exist in many other fields.

In this project, we have reduced the working model to a service that can automatically annotate occurrences of names in the scientific literature and databases. The initial target is the *International Journal of Systematic and Evolutionary Microbiology*, the publication of record for all nomenclatural changes for Bacteria and Archaea. To accomplish this objective, we have had to address several technical including transfer of the current model into a more suitable environment to simplify updating and on-the-fly generation of N4L information objects; development of tagging rules to embed links to N4L information objects into on-line content; enabling multiple resolution through the Handle server; development of mini-monographs as an improved human interface to N4L; and development of additional infrastructure to support on-the-fly translation of N4L tagged data in published content.

* Presenting author

143

GTL

The Ribosomal Database Project

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

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

Project Goals: The Ribosomal Database Project II (RDP) offers aligned and annotated rRNA sequence data and analysis service to the research community. These services help researchers with the discovery and characterization of microbes important to bioenergy production, biogeochemical cycles, and bioremediation.

Through its website (http://rdp.cme.msu.edu/), The Ribosomal Database Project II (RDP) offers aligned and annotated rRNA sequence data and analysis service to the research community (Cole et al., 2007). These services help researchers with the discovery and characterization of microbes important to bioenergy production, biogeochemical cycles, and bioremediation.

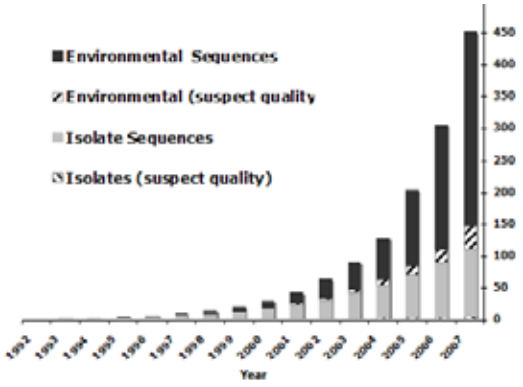


Figure 1. Increase in number of publicly available bacterial small-subunit rRNA sequences. Suspect quality sequences were flagged as anomalous by Pintail in testing with two or more reference sequences from different publications.

Updated monthly, the RDP maintains 451,545 aligned and annotated quality-controlled rRNA sequences as of November 2007 (Release 9.56; Fig. 1). All sequences are tested for sequence anomalies, including chimeras, using Pintail from the Cardiff Bioinformatics Toolkit (Ashelford et al., 2005. *Appl. Environ. Microbiol.* 71:7724-7736). The *myRDP* features introduced last year have grown to support a total of over 2400 active researchers using their *myRDP* accounts to analyze over 460,000 pre-publication sequences in 12,149 sequence library groups.

New Genome Browser: The new RDP Genome Browser allows users to browse information, including rRNA sequences, from DOE and other bacterial genome projects. The Genome Browser includes information about the organism, whether the sequenced strain is recognized as the type strain for the species, and additional links and information provided by the Genomic Standards Consortium (http://darwin.nox.ac.uk/gsc/).

New Taxomatic Visualization Tool: This interactive tool, developed under a separate GTL grant, allows researchers to choose subsets of RDP data and view a distance-based "heatmap" comparison of the user-chosen sequences. With this tool, up to two thousand sequences can be compared at one time. The researcher can interactively zoom-out on the map to gain an overview of the entire data set, or zoom-in to examine specific regions. In addition, taxonomic boundaries can be interactively displayed on the heatmap by manipulating a hierarchy of taxonomic information or by "mousing over" corresponding regions of the heatmap.

Updated Taxonomy: The RDP taxonomy has been updated to reflect changes in release 7.7 of The Taxonomic Outline of Bacteria and Archea (Garrity et al., 2007. *The Taxonomic Outline of Bacteria and Archaea. TOBA Release 7.7, March 2007.* [http://www.taxonomicoutline.org/]).

Expanded Video Tutorials: We have expanded the number of short video tutorials that demonstrate some of the more complex analytical tasks including use of *myRDP*. These tutorials average three minutes in length. They capture the screen as the tasks are performed, while the narrator explains the tasks and the choices available to the user. All tutorials are now available in Flash (with closed-captioning), Quicktime, and Windows media formats.

***Coming Soon*, RDP High-Throughput Pyrosequencing Analysis Pipeline:** New sequencing technologies, such as 454 pyrosequencing, generate tens to hundreds of thousand of partial rRNA sequences at one time and at a much lower per-sequence cost than traditional Sanger sequencing. Sequencing rRNA and other marker genes in environmental samples is a standard method of determining the bacterial composition in the sample. Conventional sequencing technologies are normally too expensive to routinely produce enough sequences to be assured of seeing any but the most abundant sequences. New sequencing technologies such as pyrosequencing have solved this problem and can produce up to hundreds

* Presenting author

of thousands of marker gene sequences from a single sample, enough to provide in-depth of analysis of bacterial composition. However most molecular ecology tools are not able to handle such large numbers of sequences. The RDP is building a Pyrosequencing Pipeline to automate the processing of these large data-sets and provide researchers with the most common ecological metrics, along with the ability to download the processed data in formats suitable for common ecological and statistical packages (Fig 2). As mentioned last year, the RDP Classifier is capable of accurately assigning such short sequences to the bacterial taxonomy (Wang et. al. 2007). In addition, we are developing new tools to align, cluster, dereplicate and simplify the compute-intensive analysis of such large sequencing libraries.

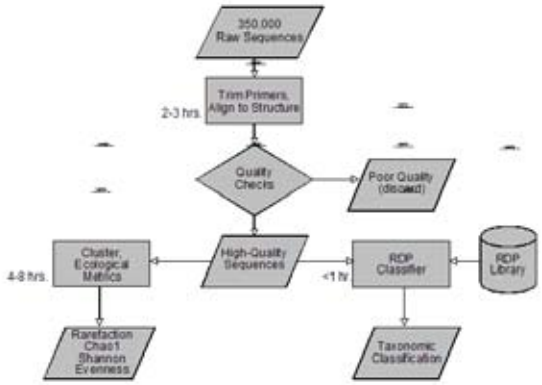


Figure 2. Flowchart showing stages in the RDP High-Throughput Pyrosequencing Analysis Pipeline being developed.

References

1. Cole, J. R., B. Chai, R. J. Farris, Q. Wang, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, A. M. Bandela, E. Cardenas, G. M. Garrity, and J. M. Tiedje. 2007. The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res.* 35 (Database issue): D169-D172; doi: 10.1093/nar/gkl889.
2. Wang, Q, G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl Environ Microbiol.* 73:5261-5267.

144

GTL

The MetaCyc and BioCyc Pathway Databases, and the Pathway Tools Software

Ron Caspi,[1] Carol Fulcher,[1] Pallavi Kaipa,[1] Markus Krummenacker,[1] Suzanne Paley,[1] Lukas Mueller,[2] Anuradha Pujar,[2] Peifen Zhang,[3] Sue Rhee,[3] and **Peter D. Karp**[1]* (pkarp@ai.sri.com)

[1]Bioinformatics Research Group, SRI International, Menlo Park, California; [2]Department of Plant Breeding and Genetics, Cornell University, Ithaca, New York; and [3]Department of Plant Biology, Carnegie Institution, Stanford, California

Project Goals: Develop the MetaCyc database describing experimentally elucidated metabolic pathways and enzymes from many organisms. Develop the BioCyc database describing metabolic pathways predicted from the complete genomes of hundreds of organisms.

Metabolic engineering demands an accurate model of the metabolic network of a target organism and the relationship of that network to the genome, plus powerful analysis tools for constructing, refining, and analyzing that model.

The MetaCyc multiorganism pathway database [1,2] describes experimentally elucidated metabolic pathways and enzymes reported in the experimental literature. MetaCyc is both an online reference source on metabolic pathways and enzymes for metabolic design, and a solid foundation of experimentally proven pathways for use in computational pathway prediction. MetaCyc version 11.6 describes 1,010 pathways from more than 1,000 organisms. The 6,500 biochemical reactions in MetaCyc reference 6,600 chemical substrates, most of which contain chemical structure information. MetaCyc describes the properties of 4,500 enzymes, such as their subunit structure, cofactors, activators, inhibitors, and in some cases their kinetic parameters. The information in MetaCyc was obtained from 15,000 research articles, and emphasizes pathways and enzymes from microbes and plants, although it also contains animal pathways.

Pathway Tools [3,4] constructs a metabolic model of an organism from its annotated genome using the following computational inference tools. The model is in the form of a Pathway/Genome Database (PGDB).

1. It predicts the metabolic pathways of the organism by recognizing known pathways from the MetaCyc database.

* Presenting author

2. It predicts which genes fill holes in those metabolic pathways (pathway holes are pathway steps for which no enzyme has been identified in the genome)
3. It predicts operons for prokaryotic genomes
4. It infers the presence of transport reactions from the names of transport proteins in the genome annotation
5. The software automatically generates a one-screen cellular overview diagram containing the metabolic and transport networks of the cell

A set of graphical editors within Pathway Tools allows scientists to refine a PGDB by adding, or modifying metabolic pathways, gene annotations, reactions, substrates, and regulatory information. The existence of an accurate knowledge base of the metabolic network is a critical resource for metabolic engineering.

The software provides a large number of operations for querying, visualization, web publishing, and analysis of PGDBs. A metabolite tracing tool supports graphical exploration of the path that a substrate follows through the metabolic network, in either the forward or backward direction. The user interactively guides the software in selecting which branches of metabolism to follow, and metabolic paths are highlighted on the cellular overview diagram.

Pathway Tools has a new emphasis on cellular regulation. It can encode information about regulation of transcription initiation and attenuation. Support for more regulatory mechanisms is under development. The Pathway Tools Regulatory Overview can display the entire transcriptional regulatory network of an organism, and can paint omics datasets onto that network to aid in their interpretation.

Pathway Tools has improved support for representing and displaying electron transport events.

Other visualization tools include automated display of metabolic pathways, reactions, enzymes, genes, and operons, and a genome browser.

SRI has applied Pathway Tools and MetaCyc to predict the pathway complements of more than 370 organisms from their complete genomes [1]. The resulting PGDBs are available through the BioCyc.org Web site. In addition, more than 75 groups outside SRI are using Pathway Tools and MetaCyc to produce PGDBs for more than 150 organisms, including the major model organisms for biomedical research (yeast, worm, fly, Dictyostelium), pathogens of biodefense interest, GTL organisms, many other bacteria and archaea, and plants (including Arabidopsis, Medicago, Rice, Tomato, and Potato).

References

1. R. Caspi et al, "The MetaCyc Database of Metabolic Pathways and Enzymes and the BioCyc Collection of Pathway/Genome Databases," *Nucleic Acids Research* in press, 2008 Database Issue.
2. P. Zhang et al, "MetaCyc and AraCyc. Metabolic Pathway Databases for Plant Research," *Plant Physiol* 138:27-37 2005.
3. P.D. Karp et al, "The Pathway Tools Software," *Bioinformatics* 18:S225-32 2002.
4. Paley, S.M. et al, "The Pathway Tools Cellular Overview Diagram and Omics Viewer," *Nucleic Acids Research* 34:3771-8, 2006.
5. Paley, S.M. et al, "Creating Fungal Pathway/Genome Databases Using Pathway Tools," *Applied Mycology and Biotechnology* 6:209-26 2006.

145

GTL

Global Credibility of Sequence Alignments

Bobbie-Jo M. Webb-Robertson[1] (bj@pnl.gov), **Lee Ann McCue**[1]* (leeann.mccue@pnl.gov), and **Charles E. Lawrence**[2] (Charles_Lawrence@brown.edu)

[1]Computational Biology and Bioinformatics, Pacific Northwest National Laboratory, Richland, Washington and [2]Division of Applied Mathematics and the Center for Computational Molecular Biology, Brown University, Providence, Rhode Island

Project Goals: The transcription regulatory network is arguably the most important foundation of cellular function, since it exerts the most fundamental control over the abundance of virtually all of a cell's functional macromolecules. The two major components of a prokaryotic cell's transcription regulation network are the transcription factors (TFs) and the transcription factor binding sites (TFBS); these components are connected by the binding of TFs to their cognate TFBS under appropriate environmental conditions. Comparative genomics has proven to be a powerful bioinformatics method with which to study transcription regulation on a genome-wide level. We will further extend comparative genomics technologies that we have introduced over the last several years, developing and applying statistical approaches to analysis of correlated sequence data (i.e. sequences from closely related

* Presenting author

species). We also plan to combine functional genomic and proteomic data with sequence data from multiple species; combining these complementary data types promises to improve our ability to predict regulatory sites of small or genus-specific regulons.

Genome sequencing initiatives have provided a wealth of bacterial sequence data, including the complete genome sequences of many closely related species like the *Shewanellas*. This rich source of data provides opportunities in comparative genomics to draw inferences on species phylogeny, metabolic capabilities, and regulation; however it also presents challenges associated with high-dimension solution space. For instance, sequence alignment is fundamental to the analysis of genome data, yet sequence alignment methods commonly focus on identifying only the best possible alignment between two sequences. However, when a single alignment is chosen for the comparison of two (or more) sequences, it is a point estimate selected from a large ensemble of all possible alignments. For example, two small sequences of length 20 generate over 2.7e+29 possible local alignments. Given the immense size of the alignment space, it is not surprising that the most probable alignments, and thus all individual alignments, often have very small probabilities.

This finding raises three questions:

1. In discrete spaces, how strongly does the available data recommend a single chosen alignment?
2. When the data provide weak evidence for any single alignment, what criteria can be used to judge the credibility, and what are reasonable limits in the degree of variation within the ensemble of alignments, that are consistent with the data?
3. How can we identify the single alignment that best represents the ensemble of alignments, and that is consistent with the data?

We present results in support of the following answers to these questions:

1. The strength of the recommendation of the data for any specific alignment is equal to its posterior probability under the assumed probabilistic model.
2. A credibility limit is the radius of the smallest hypersphere around a proposed alignment that contains a specified proportion of the posterior weighted ensemble, where the radius is measured by the number of elements by which two solutions differ. The size of this limit characterizes an alignment's credibility.
3. The alignment with the minimum credibility limit best represents the ensemble.

We find high variability in the credibility limits in the alignments of promoter sequences from closely related species. In addition, we find that the alignment with the minimum credibility limit (the ensemble centroid alignment) often differs significantly from a single "best" alignment (maximum similarity alignment). Furthermore, often the credibility limit of the maximum similarity alignment is no better than that of an alignment selected at random from the posterior weighted alignment space. We have demonstrated that credibility limits can be used to define criteria for the alignment of orthologous promoter regions from *Shewanella* species prior to motif prediction.

* Presenting author