

Systems Biology Strategies and Technologies for Understanding Microbes and Microbial Communities

Genomic and Proteomic Strategies

GTL

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

Eoin Brodie^{1*} (elbrodie@lbl.gov), Jennifer Pett-Ridge² (pettridge2@llnl.gov), Peter Weber,² Gary Andersen,¹ Meredith Blackwell,³ Nhu Nguyen,⁴ Katherine Goldfarb,¹ Stephanie Gross,³ and Paul Hoepflich² (hoepflich2@llnl.gov)

¹Lawrence Berkeley National Laboratory, Berkeley, Calif.; ²Lawrence Livermore National Laboratory, Livermore, Calif.; ³Louisiana State University, Baton Rouge, La.; and ⁴University of California, Berkeley, Calif.

Project Goals: The identification of microorganisms responsible for specific processes remains a major challenge in microbial ecology, one that requires the integration of multiple techniques. We propose to address this goal by developing a new methodology, “Chip-SIP”; combining the power of re-designed oligonucleotide microarrays with nano-scale secondary ion mass spectrometry (NanoSIMS) analyses, and linking the identity of microbes to their functional roles. Building upon the stable isotope probing approach (SIP), we will isotopically label microbial nucleic acids by growing organisms on ¹³C and ¹⁵N enriched substrates. Extracted RNA will be hybridized to a newly engineered high-density oligonucleotide microarray with a conductive surface and higher reproducibility relative to traditional microarrays. These advances in array surface chemistry will allow us to successfully analyze arrays by NanoSIMS, generating isotopic and elemental abundance images of the surface, and thereby indicating which organisms in complex consortia utilized the isotopically labeled substrate. Our first application of this new method will assign function to complex microbial community members dwelling in the hindgut of the wood-eating passalid beetle, *Odontotaenius disjunctus*. This microbial community represents a naturally-selected highly-efficient lignocellulose degrading consortium. Understanding the microbial processes by which wood-ingesting insects derive energy may aid large-scale conversion of lignocellulosic biomass into biofuels.

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.

This concept involves labeling of microbial nucleic acids following incubation with a stable isotope-labeled compound (e.g. ¹³C-cellulose or ¹⁵N₂). 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, a cyclo-olefin co-polymer plastic (COP) was identified that meets our requirements for these new microarrays (opacity comparable to glass, minimal autofluorescence, adequate hardness and temperature stability to enable surface coating processes). These COP slides were coated with ~400 angstroms ITO (indium tin oxide) and the surfaces functionalized with alkyl phosphonates. We synthesized highly reproducible oligonucleotide probe features on these alkyl phosphonate-ITO surfaces using our NimbleGen microarray synthesizer unit and successfully hybridized DNA and RNA. These alkyl phosphonate-ITO surfaces demonstrated 10X lower background and 10X higher signal:noise compared to traditional arrays. Furthermore these new ITO arrays (unlike silane arrays) may be repeatedly hybridized/stripped while maintaining high performance, allowing reuse and reduced costs (patent application filed). In preliminary NanoSIMS analyses of the ITO array surfaces, we have demonstrated the ability to analyze individual oligonucleotide probe features. We have improved environmental RNA extraction and biotin end-labeling procedures and calibrated hybridization protocols for ¹³C-labeled RNA from pure cultures and mixed bacterial consortia. Incubations with our model organism, the wood-eating passalid beetle, *Odontotaenius disjunctus*, are ongoing and include feeding experiments with ¹³C-labeled cellulose and lignin and incubation under an atmosphere of ¹⁵N₂ to identify the location and identity of lignocellulose-degrading and nitrogen-fixing bacteria respectively within this beetle hindgut.

NanoSIP: Functional Analysis of Phototrophic Microbial Mat Community Members Using High-Resolution Secondary Ion Mass Spectrometry

Luke Burow,^{1,2*} Dagmar Woebken,^{1,2*} Lee Prufert-Bebout,² Brad Bebout,² Tori Hoehler,² Jennifer Pett-Ridge,³ Steven W. Singer,³ Alfred M. Spormann,¹ and **Peter K. Weber**³ (weber21@llnl.gov)

¹Dept. of Chemical Engineering and Civil and Environmental Engineering, Stanford University, Stanford, Calif.; ²NASA Ames Research Center, Moffett Field, Calif.; and ³Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, Calif.

We are developing a new technique—nanoSIP—to measure nutrient uptake and assimilation at the single cell level. The method combines *in situ* phylogenetic and immuno-labeling methods with stable isotope probing (SIP) and nanometer-scale secondary ion mass spectrometry (NanoSIMS) analysis to link microbial identity to function in complex microbial communities. We are using this new method to study the hydrogen ecology of phototrophic microbial mats. Our studies seek to document the ecological factors that affect the pathways and efficiency with which solar energy is captured, stored in chemical form, and either dissipated within the mat or released as a usable resource. In this work, we demonstrate the application of this method to study key primary producers in a layered, dihydrogen (H₂)-evolving, phototrophic microbial mat from Elkhorn Slough, CA.

Life in phototrophic microbial mats is governed by the diel cycle. In current studies we are investigating the biogeochemical parameters of the Elkhorn Slough microbial mat during a diel cycle to learn more about metabolic processes occurring in this mat, which are responsible for H₂ production. Distinctly contrasting biogeochemistry can be observed within the surface H₂ producing layer over the cycle, with high fluxes of H₂ occurring in the dark anoxic period. H₂ flux is thought to be driven by oxygen-sensitive fermentation and N₂ fixation. These processes cease during the day due to photosynthesis, which generates oxic conditions in the mat. During the diel experiment, nitrogen fixation was measured via the acetylene reduction assay, oxygen profiles were obtained using microelectrodes and hydrogen concentrations were measured using a flux chamber. Additionally we measured photosystem II efficiency using pulsed amplitude modulated fluorometry. Mats were incubated with H¹³CO₃⁻ and ¹⁵N₂ to isotopically label newly fixed C and N (Fig. 1). We are using nanoSIP and Catalyzed Reporter Deposition-Fluorescence In Situ Hybridization (CARD-FISH) to identify the microorganisms that fix C and N and quantify their metabolic activity and thus their importance to primary production in microbial mats. Our long-term goal is to link these parameters to the investigation of diel

metabolic dynamics of microorganisms within the mats by using a global metatranscriptomic strategy.



Figure 1. Light micrograph and nanoSIP images of *Lyngbya* spp. in a microbial mat showing net carbon and nitrogen uptake. *Lyngbya* is numerically abundant in the Elkhorn microbial mat and these data show that it is an active C and N fixer in this mat. By contrast, another numerically important cyanobacteria, *Microcoleus* (not shown), fixes significant levels of C but not N in this mat.

NanoSIP: Linking Microbial Phylogeny to Metabolic Activity at the Single Cell Level Using Element Labeling and NanoSIMS Detection

J. Pett-Ridge^{1*} (pettridge2@llnl.gov), S.W. Singer,^{1*} S. Behrens,² T. Lösekann,³ W.-O. Ng,² Dagmar Woebken,² Rick Webb,⁴ B. Stevenson,⁵ D. Relman,³ A.M. Spormann,² and **P.K. Weber**¹ (weber21@llnl.gov)

¹Lawrence Livermore National Laboratory, Livermore, Calif.; ²Dept. of Civil and Environmental Engineering and ³Depts. of Microbiology & Immunology and of Medicine, Stanford University, Stanford, Calif.; ⁴Centre for Microscopy and Microanalysis, University of Queensland, Brisbane, Australia; and ⁵Dept. of Botany and Microbiology, University of Oklahoma, Norman, Okla.

We are developing a new technique—nanoSIP—to measure nutrient uptake and assimilation at the single cell level. The method combines *in situ* phylogenetic and immuno-labeling methods with stable isotope probing (SIP) and nanometer-scale secondary ion mass spectrometry (NanoSIMS) analysis to link microbial identity to function in complex microbial communities. We are using this new method to study the hydrogen ecology of phototrophic microbial mats. Our studies seek to document the ecological factors that affect the pathways and efficiency with which solar energy is captured, stored in chemical form, and either dissipated within the mat or released as a usable resource.

One of the methods we are developing uses rRNA-based fluorescence *in situ* hybridization combined with an elemental label (EL-FISH). This approach allows simultaneous phylogenetic and SIP imaging in the NanoSIMS. Fluorine or bromine atoms were introduced into cells via 16S rRNA-targeted probes, which enabled phylogenetic identification of individual cells by NanoSIMS elemental imaging. To overcome the natural halogen backgrounds, we used the catalyzed reporter deposition (CARD)-FISH technique

with halogen-containing fluorescently labeled tyramides as substrates for the enzymatic tyramide deposition. The relative cellular abundance of fluorine or bromine after EL-FISH exceeded natural background concentrations by up to 180-fold, and allowed us to distinguish target from non-target cells in NanoSIMS fluorine or bromine images. The method was optimized on axenic cultures and applied to a dual-species consortium (filamentous cyanobacterium and heterotrophic alpha-proteobacterium) and complex microbial aggregates from human oral biofilms. We have also conducted a multi-factorial experiment to test the effects of the EL-FISH technique on isotopic enrichment in both pure and mixed cell cultures. We found significant effects of fixation reagent and CARD-FISH that must be accounted for when interpreting isotopic enrichment data.

The other visualization method we are developing is based on immuno-labeling of functional proteins with nanoparticles. The nanoparticles can be imaged directly with the NanoSIMS to enable simultaneous imaging of functional protein and SIP imaging. The advantage of this approach over correlated TEM-NanoSIMS studies is that it eliminates the TEM step, it allows for thick-section analysis, and electron dense phases in the sample are readily distinguished by elemental composition in the NanoSIMS. We have succeeded in imaging Au nanoparticles in the NanoSIMS at standard antibody densities. We are working on developing this method in axenic *Lyngbya* cultures and complex microbial communities, focusing on nitrogen fixation in cyanobacteria.

—
GTL

Production of Extracellular Polymeric Substances by a Natural Acidophilic Biofilm

Y. Jiao,^{1*} (jiao1@llnl.gov), M. Shrenk,² G. Cody,² J.F. Banfield,³ and M.P. Thelen¹

¹Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, Calif.;

²Geophysical Laboratory, Carnegie Institution of Washington, Washington, D.C.; and ³Dept. of Environmental Science, Policy, and Management, University of California, Berkeley, Calif.

Project Goals: This is part of a project led by JF Banfield at the University of California at Berkeley. In this subproject (MPT, PI), our goals are to identify and characterize the functions of biofilm constituents isolated from a low-diversity community of microbes engaged in iron oxidation. Biofilms are collected from an EPA Superfund site characterized by extremes of high heat and acid pH, leading to acid mine drainage pollution. Both proteins and extracellular polymeric substances (EPS) from biofilms at different growth stages are under current investigation, towards a comprehensive understanding of the ecology and organization of this acidophilic community.

Biofilms are ubiquitous in nature. In most cases, biofilms facilitate the formation of specialized microbial communities adhering to natural or artificial surfaces through the production of extracellular polymeric substances (EPS). The embedding of microorganisms within complex polysaccharides and other biopolymers promotes microbial assemblages, cell adhesion, and community protection from hostile environmental conditions. As major structural components, EPS provides spatial organization and structural stability to the microbial community. The exact composition of EPS varies substantially between different biofilms, and it remains a considerable challenge to provide a complete biochemical profile for interpreting how changes in EPS constituents affect community organization and development. Here we present a framework of EPS analysis whereby the correlation between EPS composition and potential microbial functions can be explored.

As part of our ongoing proteogenomic investigation of acidophilic communities collected in an acid mine drainage site (Richmond Mine, Iron Mountain, California; see Mueller et al. poster), we examined the properties of EPS as a function of biofilm development. EPS was extracted by ethanol precipitation after biofilms were homogenized in a sulfuric acid solution. The quality and quantity of EPS obtained from early growth stage (GS1) and mid-stage (GS2) biofilms were evaluated. More than twice the amount of EPS was obtained from GS2 compared to that of GS1, with approximately 15 and 40 mg of EPS per gram of dry weight for GS1 and GS2, respectively. Chemical composition analysis indicated the presence of carbohydrate and heavy metals, and minor quantities of protein and DNA, although the relative concentration of each component varied between the two EPS samples. EPS from the GS2 biofilm contains significantly higher concentrations of carbohydrate and heavy metals compared to GS1. Glycosyl composition analysis indicates that both EPS samples are composed primarily of galactose, glucose, heptose, rhamnose and mannose, and the relative amount of individual sugars varies substantially with developmental stage (see Table 1).

Table 1. Glycosyl Composition Analysis of GS1 and GS2 samples

Glycosyl residue	GS1 (Mol%)	GS2 (Mol%)
Galactose (Gal)	51.9	18.9
Glucose (Glc)	21.2	16.7
Heptose (Hep)	12.2	33.0
Rhamnose (Rha)	11.1	13.3
Mannose (Man)	8.7	8.0
Xylose (Xyl)	n.d.	2.6
3 Deoxy-2-manno-2 Octulosonic acid (KDO)	n.d.	7.8
3OH C16 Fatty Acid	n.d.	+
N Acetyl Glucosamine (GlcNAc)	n.d.	+

Table 2. Comparison of glycosyl linkage analysis of GS1 and GS2 samples.

Glycosyl Residue	GS1 (mol%)	GS2 (mol%)
terminally linked hexofuranosyl residue (t-hexf)	21.2	10.9
4-linked glucopyranosyl residue (4-Glc)	13.1	8.2
terminally linked rhamnopyranosyl residue (t-Rha)	12.2	15.5
terminally linked heptopyranosyl residue (t-Hep p)	9.9	10.6
3,4-linked galactopyranosyl residue (3,4-Gal)	8.6	3.7
terminally linked mannopyranosyl residue (t-Man)	7.2	12.3
2-linked mannopyranosyl residue (2-Man)	4.6	9.6

Additionally, carbohydrate-linkage analysis reveals multiply linked heptose, galactose, glucose, mannose and rhamnose (see Table 2), similar to the complex, branched polysaccharides found in plant cell walls, and perhaps also expected in an environmental biofilm matrix. Interestingly, much of the glucose measured is in the 4-linked form. Consistently, solid-state NMR analysis of EPS samples indicates that up to 25% of the EPS is cellulose ($\beta 1 \rightarrow 4$ glucan).

Besides providing a structural element in biofilms, EPS in acid mine drainage microbial communities may offer protection from toxic heavy metals through diffusion limitation, and facilitate nutrient flux between different members within the community by acting as a carbon source/sink. In addressing the covariance pattern of EPS composition with biofilm developmental stages and microbial processes, we hope to gain a fundamental understanding of how EPS functions in natural microbial communities.

This work was funded by the DOE Genomics:GTL Program grant number DE-FG02-05ER64134, and performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

—
GTL

Soil Community Metagenomics at the DOE's Climate Change Research Sites

Cheryl Kuske^{1*} (kuske@lanl.gov) and the FACE soil metagenomics working group: Gary Xie,¹ John Dunbar,¹ Larry Ticknor,¹ Yvonne Rogers,¹ Shannon Silva,¹ La Verne Gallegos-Garcia,¹ Stephanie Eichorst,¹ Shannon Johnson,¹ Don Zak,² Rytas Vilgalys,³ Chris Schadt,⁴ Dave Evans,⁵ Patrick Magonigal,⁶ Bruce Hungate,⁷ Rob Jackson,³ David Bruce,^{1,8} and Susannah Tringe⁸

¹Los Alamos National Laboratory, Los Alamos, N.M.; ²University of Michigan, Aspen FACE site, Ann Arbor, Mich.; ³Duke University, Pine FACE site, Durham, N.C.;

⁴Oak Ridge National Laboratory, Sweetgum FACE site, Oak Ridge, Tenn.; ⁵Western Washington University, Desert FACE site, Bellingham, Wash.; ⁶Smithsonian Institution, Estuary FACE site, Washington, D.C.; ⁷Northern Arizona University, Scrub Oak OTC site, Flagstaff, Ariz.; and ⁸DOE Joint Genome Institute, Walnut Creek, Calif.

Project Goals: Our overarching vision is to establish a science strategy and implementation pipelines to obtain a comprehensive, field-scale understanding of the structure of soil microbial communities and their functional processes that play critical roles in terrestrial carbon sequestration and response to climate change perturbations. Our three goals provide complementary science and technology information to achieve this overall vision. (1) Understand the impacts of long-term elevated CO₂ and other environmental factors (ozone, nitrogen interactions) on the structure and activities of soil microbial communities, at the DOE's six Free Air CO₂ Enrichment (FACE) and Open Top Chamber (OTC) experimental field sites. (2) Establish a multi-tier genomics-based analysis and ecological integration capability that links DOE JGI sequencing technology with need to understand functional abilities of soil microbial communities in an ecological setting, and to provide a platform that allows comparison of those processes across different terrestrial ecosystems. (3) Improve our basis for comparison of soil populations involved in carbon cycling and climate change response by expanding the functional genes and proteins we can use to detect and monitor these populations. Metagenomics can be usefully coupled with large, field-scale ecological studies that are focused on quantifying and modeling key processes.

In the next two years, six of the DOE's long-term free air CO₂ enrichment (FACE) and open top chamber (OTC) experiments will come to completion, offering an excellent opportunity to determine the effects of over ten year of elevated CO₂ treatment on below-ground ecosystem processes and the soil microbial communities responsible for those processes. The long-term FACE and OTC experiments encompass forest, scrubland, desert, and wetlands, allowing comparison of belowground responses of very different terrestrial ecosystems to elevated CO₂. Soil microbiota play critical roles in cycling carbon and nitrogen in terrestrial ecosystems, and their contributions have local, regional and global impacts on terrestrial carbon storage and cycling.

A multi-institutional, FACE soil metagenomics working group has been assembled to study soil microbial community structure across the FACE and OTC sites. Using a variety of comparative metagenomic sequencing approaches, we are addressing two questions: (a) Has long-term elevated CO₂ treatment affected the abundance and composition of soil microbiota in the soil? (b) Have the soil communities in different ecosystems responded similarly or in different ways to elevated CO₂? Three targeted metagenomic strategies are currently being used to address these questions across the DOE's FACE and OTC sites: (a) taxonomic profiling of the bacteria, archaea, and fungi in soils exposed to elevated or ambient CO₂ conditions, (b) sequencing suites of functional genes that are important in carbon and nitrogen cycling,

(c) sequencing the soil fungal transcriptome in sites where we have evidence of the importance of fungal activities in response to elevated CO₂. We plan to also investigate seasonal responses to elevated CO₂, and the interactive effects of ozone and addition of soil nitrogen using these approaches. We will include a total community metagenomics approach in our future studies. We present here the initial results of rRNA profiling of bacterial and fungal communities across treatments at each of the FACE and OTC sites, and of high depth coverage of fungal rRNAs using 454 pyrosequencing for the aspen FACE site.

GTL

Flow Sorting and Whole Genome Amplification of Individual Microbes

Sébastien Rodrigue^{1*} (s_rod@mit.edu), Rex R. Malmstrom,^{1*} Aaron Berlin,² Matthew Henn,² and Sallie W. Chisholm^{1,3} (chisholm@mit.edu)

¹Dept. of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Mass.; ²Broad Institute, Cambridge, Mass.; and ³Dept. of Biology, Massachusetts Institute of Technology, Cambridge, Mass.

Project Goals: Develop a high throughput, ultra-clean pipeline for performing whole genome amplification on individual marine bacteria using flow cytometric cell sorting; determine the best technologies and strategies to sequence genome from individual microbial cells.

Genome sequencing of single cells is becoming a reality through the combination of whole genome amplification (WGA) and next-generation sequencing technologies. We have developed a high throughput, ultra-clean pipeline for performing WGA on individual marine bacteria using flow cytometric cell sorting and automated liquid handlers. The pipeline has already enabled WGA of several hundred individual bacterial during a single session. To validate the pipeline for sequencing, we performed WGA on individual cells from a culture of *Prochlorococcus* strain MED4, for which a reference genome had previously been assembled by Sanger sequencing. The resulting DNA was prepared for sequencing on the Roche 454-FLX and Illumina platforms. Both platforms yielded comparable coverage results, although 454 assemblies were superior. With 454-FLX, ~98% of the reference genome was covered at >1X after one sequencing run of single cell amplified DNA. A *de novo* assembly without a reference genome resulted in 523 contigs covering 84% of the reference *Prochlorococcus* MED4 genome, while unamplified DNA extracted from a culture containing billions of cells was assembled into only 7 contigs covering 99.85% of the reference genome. This difference between a single cell and billions of cells is due to uneven amplification of DNA from an individual cell; some genomic regions had >1000X coverage while others <5X. A comparison of four identically amplified *Prochlorococcus* MED4 single cell genomes revealed no patterns in amplification bias, in agreement with previous reports suggesting that uneven coverage is a stochastic

process. To mitigate this uneven coverage, we developed a normalization procedure for 454 sequencing libraries that preferentially removes highly abundant sequences. This normalization procedure can reduce overall coverage variation by two orders of magnitude and dramatically improves *de novo* assembly of amplified single cell genomes.

GTL

Viruses Hijacking Cyanobacterial Carbon Metabolism

Luke R. Thompson^{1*} (luket@mit.edu), JoAnne Stubbe,^{1,2} and Sallie W. Chisholm^{1,3}

¹Depts. of Biology, ²Chemistry, and ³Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Mass.

<http://chisholmlab.mit.edu/>

Project goals: Characterize the cellular machinery of *Prochlorococcus* and its phage, as a model system for photosynthetic energy conversion.

Cyanophage infecting the marine cyanobacteria *Prochlorococcus* and *Synechococcus* carry several genes involved in the pentose phosphate pathway (PPP), a nighttime alternative to the Calvin cycle that generates NADPH and ribose. Oxidizing conditions in cyanobacteria at night, when photosystem I cannot generate NADPH, favor flux through the PPP, including the key enzyme transaldolase. Many cyanophage carry a transaldolase gene (*talC*) that differs markedly in structure from the host transaldolase (*talA*), suggesting that its acquisition and maintenance by cyanophage stems from functional differences with the host transaldolase. We have shown that the host enzyme is subject to oxidation in aerobic conditions, requiring reductant for full activity, whereas the phage transaldolase has no such effect. Site-directed mutagenesis of host transaldolase cysteines suggests that a mechanism independent of disulfide bond formation is responsible for this redox effect. We have recently found that another PPP-related gene, for the photosynthetic regulatory protein CP12, is also carried by many cyanophage. In cyanobacteria and other phototrophs, CP12 binds and deactivates two Calvin cycle enzymes under nighttime oxidizing conditions, promoting flux through the PPP. It therefore seems that cyanophage promote flux through the PPP by encoding not only enzymes but also a regulatory protein that inhibits the competing Calvin cycle. Phage infection of some cyanobacteria is known to lead to oxidizing conditions. Oxidative inactivation of host transaldolase and activation of phage CP12 may therefore be physiologically important, allowing cyanophage to produce NADPH and ribose for nucleotide biosynthesis and genome replication. Abundance patterns in phage genomes of *talC*, *cp12*, and two phage-encoded PPP dehydrogenases (*zwf* and *gnd*) are mirrored in environmental sequence databases, suggesting that the metabolic hijacking of cyanobacteria by cyanophage may be a globally important phenomenon.

Quantitative Proteomics of *Prochlorococcus*: Towards an Integrated View of Gene Expression and Cellular Stoichiometry

Jacob R. Waldbauer^{1*} (jwal@mit.edu) and **Sallie W. Chisholm**² (chisholm@mit.edu)

¹Joint Program in Chemical Oceanography, Massachusetts Institute of Technology and Woods Hole Oceanographic Institution, Cambridge, Mass.; ²Dept. of Civil & Environmental Engineering and Dept. of Biology, Massachusetts Institute of Technology, Cambridge, Mass.

The marine cyanobacterium *Prochlorococcus* is the most abundant oxygenic photosynthetic organism on earth and a key component of oceanic carbon and nutrient cycling. With 13 sequenced genomes of cultured isolates and a wealth of observational data from natural environments, it is also a prime model system for understanding the molecular bases of microbial ecology and biogeochemistry. The small sizes of both genomes (~1700 genes) and cells (~0.6 micron) of *Prochlorococcus* also make tractable a relatively complete inventory of gene products and systems biology analysis of its streamlined metabolism. We have implemented a metabolic ¹⁵N-labeling strategy to perform quantitative proteomic measurements, and find that it is comparable in accuracy and precision to mRNA-level measures of gene expression, including RT-qPCR, microarrays and RNA-sequencing.

In the oceans, cell division in *Prochlorococcus* populations is well-synchronized to the diel light-dark cycle, a phenomenon that can be reproduced in laboratory culture. Microarray experiments have demonstrated that most of the genome shows a substantial cyclicality of expression over this diel cell cycle. These results point to a 'just-in-time' mode of metabolism, where limited resources are apportioned in turn to various cellular processes in a coordinated fashion. We are extending these observations with quantitative proteomics, to explore differences in timing and magnitude between mRNA- and protein-level gene expression. The ¹⁵N metabolic labeling strategy, in combination with accurate cell counts by flow cytometry, affords the ability to quantify proteins on a per-cell basis. Our aim is to develop a comprehensive picture of gene product stoichiometries (including both transcript and protein abundances) over the diel cell-division cycle of *Prochlorococcus*, which will help elucidate the molecular processes underlying microbial growth and carbon fixation in the global ocean.

Bacterioplankton Community Transcriptional Response to Environmental Perturbations

Jay McCarren* (mccarren@mit.edu), Rex Malmstrom, Yanmei Shi, Sebastien Rodrigue, **Sallie W. Chisholm**, and **Edward DeLong**

Dept. of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Mass.

Environmental metagenomic techniques have revealed the vast genetic potential present in natural microbial communities. A remaining challenge is to examine gene expression in response to environmental variation in complex communities, with one goal being to determine the kinetics and impact of microbial response in gene expression on biogeochemical processes. Next generation sequencing has enabled cDNA profiling of microbial communities and we are developing this approach to track community responses to various environmental perturbations. Using this approach, we conducted field experiments to identify the genes, pathways, and organisms that respond to enrichments of either carbon substrates or nutrients in open ocean planktonic microbial communities. By tracking changes in whole community and *Prochlorococcus*-specific DNA and cDNA abundance throughout the experiment, we aim to identify gene suites that are up or down-regulated in response to these nutrient perturbations. Such investigations will also reveal which taxonomic components of the community respond to these amendments. Determining which genes allow which organisms to access new sources of nutrients and/or organic carbon substrates will shed light on carbon cycling dynamics in the environment. These transcriptomic experiments are providing new and specific insight into the functional dynamics and metabolic responses of natural microbial communities to environmental perturbation.

We are focusing both on whole community analyses (synecology), as well as on specific groups like *Prochlorococcus* (autecology). The autecological approach with *Prochlorococcus* has the advantage that these cyanobacteria can be separated from the bulk community by flow cytometric sorting. Using this approach, we conducted field experiments to identify the genes, pathways, and organisms that respond to enrichments of carbon substrates (high molecular weight dissolved organic material; HMW-DOM), and nutrients (inorganic PO₄ and mesopelagic water amendments) in open ocean planktonic microbial communities. By tracking changes in whole community and *Prochlorococcus*-specific DNA and cDNA abundance throughout the experiment, we have identified specific gene suites that are up or down-regulated in response to these nutrient perturbations. Phylogenetic analyses based on both community DNA and RNA abundances revealed that specific components of the bacterial community responded differentially to HMW-DOM additions. Gammaproteobacterial orders *Alteromonadales*, *Thiotrichales* and *Xanthomonadales* responded positively to this DOM amendment, whereas *Prochlorococcus* and *Pelagibacter* relative cell numbers decreased. Preliminary

transcriptome analyses indicate that particular functional gene suites show parallel, functionally variable expression patterns in response to HMW-DOM enrichments. Detailed analyses of these transcriptomic experiments are providing new and specific insight into the functional dynamics and metabolic responses of natural microbial communities to environmental perturbation.

GTL

Grand Challenge in Membrane Biology: A Systems Biology Study of the Unicellular Diazotrophic Cyanobacterium *Cyanothece* sp. ATCC 51142

Michelle Liberton^{1*} (miliberton@biology2.wustl.edu), Jana Stöckel¹, Alice C. Dohnalkova², Galya Orr³, Jon M. Jacobs³, Thanura Elvitigala¹, Eric A. Welsh¹, Hongtao Min⁴, Jörg Toepel⁴, Thomas O. Metz³, Hans Scholten³, Michael A. Kennedy³, Garry W. Buchko³, Nicole M. Koropatkin⁵, Rajeev Aurora⁶, Bijoy K. Ghosh¹, Teruo Ogawa⁷, Jason E. McDermott³, Katrina M. Waters³, Christopher Oehmen³, Gordon A. Anderson³, Thomas J. Smith⁵, Richard D. Smith³, Louis A. Sherman⁴, David W. Koppenaal^{2,3}, and **Himadri B. Pakrasi**¹ (pakrasi@wustl.edu)

¹Dept. of Biology, Washington University, St. Louis, Mo.; ²Environmental Molecular Sciences Laboratory, ³Fundamental & Computational Sciences Directorate, Pacific Northwest National Laboratory, Richland, Wash.; ⁴Dept. of Biological Sciences, Purdue University, West Lafayette, Ind.; ⁵Donald Danforth Plant Science Center, St. Louis, Mo.; ⁶Dept. of Molecular Microbiology and Immunology, Saint Louis University School of Medicine, St. Louis, Mo.; and ⁷National Laboratory of Plant Molecular Genetics, Institute of Plant Physiology and Ecology, Shanghai Institute for Biological Sciences, Shanghai, China

Project Goals: Membrane processes are critical to solving complex problems related to energy production, carbon sequestration, bioremediation, and other issues in energy and environmental science. Understanding how membrane processes fit into the overall cellular physiology and ecology requires a systems-level analysis of the genetics, biochemistry, and biophysics of membrane components and how molecular machines assemble, function, and disassemble as a function of time. *Cyanothece* is a marine cyanobacterium capable of oxygenic photosynthesis, nitrogen fixation, and heterotrophic growth in the dark. This unicellular organism has evolved an elaborate diurnal rhythm to temporally separate nitrogen fixation from oxygen production during oxygenic photosynthesis because of the oxygen-sensitive nature of the nitrogenase enzyme. The diurnal patterns of nitrogen fixation in the dark and photosynthesis in the light make *Cyanothece* a unique model organism for studying solar energy

harvesting, carbon sequestration, metal acquisition, and hydrogen production. We sought to use systems biology approaches to determine the underlying cell signaling networks that govern the functions of cyanobacterial membranes and their components to accomplish this dramatic diurnal cycling. Our research brings together expertise in microbiology, biochemistry, proteomics and metabolomics, structural biology, imaging, and computational modeling and bioinformatics to achieve these objectives.

Cyanothece 51142 is a marine cyanobacterial strain notable for its ability to perform oxygenic photosynthesis and nitrogen fixation in the same single cell. These incompatible processes are temporally separated: photosynthesis is performed during the day and nitrogen fixation at night. As part of a complex diurnal cycle, these cells accumulate and subsequently mobilize storage inclusion bodies, specifically glycogen (carbon) and cyanophycin (nitrogen), making them natural biological batteries. In order to understand at a systems level how *Cyanothece* accomplishes these complex metabolic processes, we have undertaken a combination of ultrastructural, physiological, genomic, transcriptomic, proteomic, and metabolomic studies of this organism. High-resolution 3-D electron microscopy revealed that *Cyanothece* cells have a single extensive internal thylakoid membrane system. The genome of *Cyanothece* was sequenced (1) and found to contain a unique arrangement of one large circular chromosome, four small plasmids, and one linear chromosome. Global transcriptional analyses (2) uncovered 30% of genes with cyclic expression patterns and pinpointed a significant impact of nitrogen fixation on the diurnal cycle of different fundamental pathways. We have utilized the high-throughput accurate mass and time (AMT) tag approach to examine the proteome of *Cyanothece* 51142, and identified a total of 3,470 proteins with high confidence, which is approximately 65% of the predicted proteins based on the completely sequenced genome. These studies, as well as metabolite profiling, structural studies (3,4), and physiological measurements, when coupled with computational analysis and metabolic modeling, describe an organism in which tight control of cellular processes linked to storage of metabolic products for later usage is paramount for ecological success.

References

1. Welsh EA, *et al.* (2008) The genome of *Cyanothece* 51142, a unicellular diazotrophic cyanobacterium important in the marine nitrogen cycle. *Proc Natl Acad Sci U S A* 105: 15094-15099.
2. Stöckel J, *et al.* (2008) Global transcriptomic analysis of *Cyanothece* 51142 reveals robust diurnal oscillation of central metabolic processes. *Proc Natl Acad Sci U S A* 105: 6156-6161.
3. Buchko GW, *et al.* (2008) Insights into the structural variation between pentapeptide repeat proteins - Crystal structure of Rfr23 from *Cyanothece* 51142. *Journal of Structural Biology* 162(1): 184-192.
4. Koropatkin N, *et al.* (2007) The structure of a cyanobacterial bicarbonate transport protein, CmpA. *Journal of Biological Chemistry* 282(4): 2606-2614.

This work is part of a Membrane Biology EMSL Scientific Grand Challenge project at the W. R. Wiley Environmental Molecular Science Laboratory, a national scientific user facility sponsored by the U.S. Department of Energy's Office of Biological and Environmental Research (BER) program located at Pacific Northwest National Laboratory. PNNL is operated for the Department of Energy by Battelle.

GTL

Functional Analysis of Trace Nutrient Homeostasis in *Chlamydomonas* using Next Generation Sequencers

Sabeeha Merchant^{1,2} (sabeeha@chem.ucla.edu), Madeli Castruita,² David Casero,³ Janette Kropat,² Steven Karpowicz,² Shawn Cokus,³ and Matteo Pellegrini^{1,3*}

¹The Institute for Genomics and Proteomics, ²Dept. of Chemistry and Biochemistry, and ³Dept. of Molecular, Cell and Developmental Biology, University of California, Los Angeles, Calif.

Project Goals: Our goal is to use next generation sequencing technology to annotate algal genomes with the goal of increasing our understanding of algal biology.

Chlamydomonas, a chlorophyte alga in the green plant lineage, is a choice model organism for the study of chloroplast-based photosynthesis and cilia-based motility. The 121 Mb draft genome sequence, determined at 13X coverage is estimated to encode approximately 15,000 protein coding genes. Besides the pathways for oxygen evolving photosynthesis, dark respiration of acetate and hydrogen production, the gene repertoire reveals less-studied pathways for fermentative metabolism, suggestive of extraordinary metabolic flexibility. The operation of these bioenergetic pathways is dependent on metal cofactors like copper, iron, manganese and zinc, and accordingly these elements are essential nutrients for *Chlamydomonas*. In a copper-deficient environment, *Chlamydomonas* will modify the photosynthetic apparatus by substituting a heme protein—Cyt *c*₆—for an abundant copper protein—plastocyanin—that accounts for about half of the intracellular copper. This modification is viewed as a copper sparing mechanism and is dependent on a plant specific transcription factor CRR1. We have used digital gene expression (DGE) and RNA-Seq methodology to characterize the *Chlamydomonas* transcriptome under steady conditions of various degrees of copper-deficiency and in a bloom situation where cells deplete the copper as they divide. Both methods are quantitative and show excellent correlation with real time PCR indicative of a large dynamic range relative to microarrays. Direct vs. indirect responses to copper-deficiency are distinguished by comparison of the *crr1* transcriptome to that of wild-type cells. The analyses indicate previously unknown modifications of the photosynthetic apparatus and the potential for modification of bioenergetic pathways.

Whole-Genome Comparative Analyses Across an Environmental Gradient Reveal Surprisingly Rapid Bacterial Adaptation Mediated by Horizontal Gene Transfer

Alejandro Caro,¹ Jie Deng,² Jennifer Auchtung,² James Tiedje,² and Konstantinos Konstantinidis^{1*} (kostas@ce.gatech.edu)

¹School of Civil and Environmental Engineering and School of Biology, Georgia Institute of Technology, Atlanta, Ga.; and ²Center for Microbial Ecology, Michigan State University, East Lansing, Mich.

Project Goals: Dr. Konstantinidis' group at Georgia Tech will conduct computational analyses of genomic, transcriptomic, proteomic, metabolomic and physiological data for *Shewanella* organisms to provide insight into the evolution, speciation and traits that determine niche optimization of these organisms. The work is going to be performed exclusively in-silico (computational), thus, no human or animal subjects or recombinant DNA etc. is involved. Dr. Konstantinidis' group will work cooperatively with Michigan State University and other members of the *Shewanella* Federation to achieve these goals. Georgia Tech's work is under Objective 4, Task 1 in the above project, entitled "Genetic and ecophysiological bases defining the core and diversification of *Shewanella* species".

How fast bacteria adapt to environmental fluctuations and to degrade new xenobiotic compounds remains a poorly understood issue of paramount importance for DOE's bioremediation efforts. One way that bacteria adapt is through changes in their genomic makeup, such as those caused by the exchange of genetic material among different bacterial strains. To provide new insights into these issues, we have sequenced and compared four strains of *Shewanella baltica*. These strains originated from four different depths of the stably stratified Baltic Sea, characterized by difference redox potentials and nutrient availability. The strains showed very similar evolutionary relatedness among each other, with their average genomic nucleotide identity being ~97%. Despite their comparable relatedness, the two strains isolated from more similar depths shared significantly more genomic islands compared to strains from different depths. The islands appeared to carry the ecologically important genes that determined strain's successful adaptation to the unique characteristics of the particular depth. Remarkably, the majority of these genes and an additional ~20% of the core genes (i.e. present in all genomes) showed 99.8-100% nucleotide identity between the two strains, suggesting that they had been horizontally exchanged between the strains in very recent evolutionary time. These results were validated against a larger collection of strains from the Baltic Sea using DNA-DNA microarrays. Collectively, our findings reveal that genomic adaptation could be very rapid, especially among spatially co-occurring strains, and advance our

understanding of population adaptation at redox interfaces, the most important environments for the transfer of pollutants relevant to DOE's missions.

—
GTL

A Laboratory Scientist Encounters Genome Sequences

F. William Studier* (studier@bnl.gov)

Biology Dept., Brookhaven National Laboratory, Upton, N.Y.

Project Goals: Improve genome annotation.

A consortium of scientists from Korea, France and the USA has determined the genome sequences of two B strains of *Escherichia coli*: REL606 is a strain used by Richard Lenski and colleagues for long-term studies of evolution in the laboratory; BL21(DE3) is a strain I constructed to be a host for expressing recombinant proteins under control of T7 RNA polymerase. Both are descended from the strain Delbrück and Luria named *E. coli* B in 1942 and which was adopted as the common host for the study of phages T1 to T7. Intensive study of laboratory strains derived from *E. coli* B and the closely related *E. coli* K-12, introduced by Tatum's 1944 work on biochemical mutants, is the basis for much of our current understanding of molecular genetics.

Comparison of the genome sequences of the two B strains has revealed the effects of the several different laboratory manipulations known to have occurred in the two lineages. Also revealed was an unknown misidentification in the literature that initially confused the analysis. The poster will summarize the differences between the two genome sequences, how they came about, and what they reveal about the effects of common laboratory practices.

—
GTL

Coupling Function to Phylogeny via Single-Cell Phenotyping

Michael C. Konopka^{1*} (mkonopka@u.washington.edu), Sarah C. McQuaide,² Samuel Levine,¹ Ekaterina Latypova,³ Tim J. Strovas,⁴ Marina G. Kalyuzhnaya,³ and **Mary E. Lidstrom**^{1,3} (lidstrom@u.washington.edu)

¹Dept. of Chemical Engineering, ²Dept. of Electrical Engineering, ³Dept. of Microbiology, and ⁴Dept. of Bioengineering, University of Washington, Seattle, Wash.

Project Goals: One of the major challenges in understanding microbial community function in natural environments is linking genome-level sequences to function and role in the ecosystem. This project takes a new approach that couples a function-based live cell presorting step to

single-cell analysis at both the physiological and genetic levels, using Lake Washington sediment and C1 cycling as the model. This approach will allow a culture-independent enrichment of live cells involved in specific functions, analysis of a variety of phenotypic capabilities at the single cell level, then targeting of those cells that test positive for specific functions for further culture-dependent and sequencing analysis. In a relatively rapid and high throughput manner, this system will identify cells with functions of interest and carry out a set of phenotypic tests on those cells. Subsequently, both culture-dependent and independent methods will be used to obtain sequence and phenotypic information that will couple function to genomics. The objectives of this project are 1) develop new technology for presorting functional populations and analyze them at the single cell level for both phenotypic and genomic parameters, and 2) apply this approach to populations from Lake Washington sediments to couple functional and genomic datasets at the single cell level.

Respiration represents the largest sink of organic matter in the biosphere, and is a fundamental component of global carbon flow. For decades, the balance between respiration and production has been considered an essential characteristic of ecosystems. From an environmental viewpoint, respiration is a key for understanding ecosystem functioning, structure and dynamics. We use respiration as a main physiological parameter to assess specific functions of yet uncultivable microbial population/individual cells from natural niches. Here we present three approaches for detection and characterization of substrate-linked respiration: **1. Respiratory Detection System (RDS)** for bulk measurement of ecosystem function; **2. Respiration Response Imaging (RRI)** utilizing a fluorescent redox indicator for real-time detection of methylotrophic abilities of individual cells in a bulk system; **3. Microobservation Chamber** as a single cell analysis system for carrying out the physiological and genomic profiling of cells capable of respiring C₁-compounds.

1. Respiratory Detection System. To evaluate the potential of the RDS for environmental applications, the system was tested on Lake Washington sediment samples. For decades, this lake sediment has served as model system to study benthic carbon flux and microbial activities. It was demonstrated that aerobic reactions are restricted to a narrow zone of the lake sediment, and oxygen is depleted within the top centimeter of the sediment. We used the RDS to measure respiration rate in the top aerobic layer of the sediment cores. The minimum incubation time required for the rate calculation was 30-35 minutes (more than 40 times shorter than a routine incubation approach). The rate of oxygen consumption in the RDS system for fresh samples of the sediment was 1.4±0.05 mmol m⁻³ h⁻¹. The values obtained fall within the range previously reported for freshwater lakes (Giorgio and Williams, 2005). Overall, the RDS system provides precise measurements of oxygen decline over a short period of time (minutes, not hours), and thus shows great potential as a routine application in environmental studies.

2. Respiration Response Imaging. With RRI it is possible to detect the response of a natural population of microbial cells to an environmental change/stress at the very moment that it takes place at a single cell level. RedoxSensor green is a novel fluorogenic redox indicator dye (Molecular Probes, Invitrogen) that yields green fluorescence (488 nm excitation) after modification by bacterial reductases. Since respiration rate is proportional to fluorescence intensity, response to a new substrate or environmental change is easily detectable for individual cells from the entire population. We demonstrated using RRI to perform high throughput analysis of responses for environmental samples.

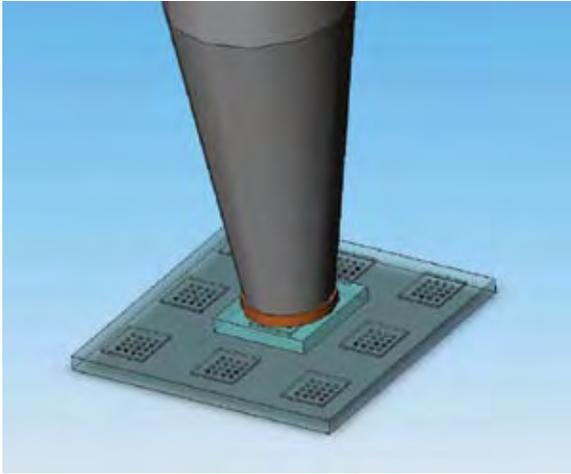


Figure 1. Glass chip containing arrays of microwells with piston sealing center array.

3. Microobservation Chamber. This system was based on an existing Microscale Life Sciences Center microwell-based platform consisting of a glass chip fabricated in-house with nine 4x4 arrays of etched microwells. Each microwell contains Pt-Porphyrin polystyrene beads as an oxygen sensor. The phosphorescent lifetime of the porphyrin sensor is inversely related to the oxygen concentration inside the well. Cells are seeded and the chip is placed inside a macro-well containing additional medium that bathes the chip. During an experiment, a piston with a flat glass tip is brought down over the top of the glass wells, coming into a pressure contact with the raised platform and diffusionally sealing each well from its neighbors (Fig. 1). Once the wells are sealed, lifetime measurements are made of each individual well in the array simultaneously and data are processed in real-time so that the oxygen consumption in each well can be monitored. After the oxygen consumption rate is detected inside the individual wells of an array, the piston is brought up and the wells are allowed to reoxygenate.

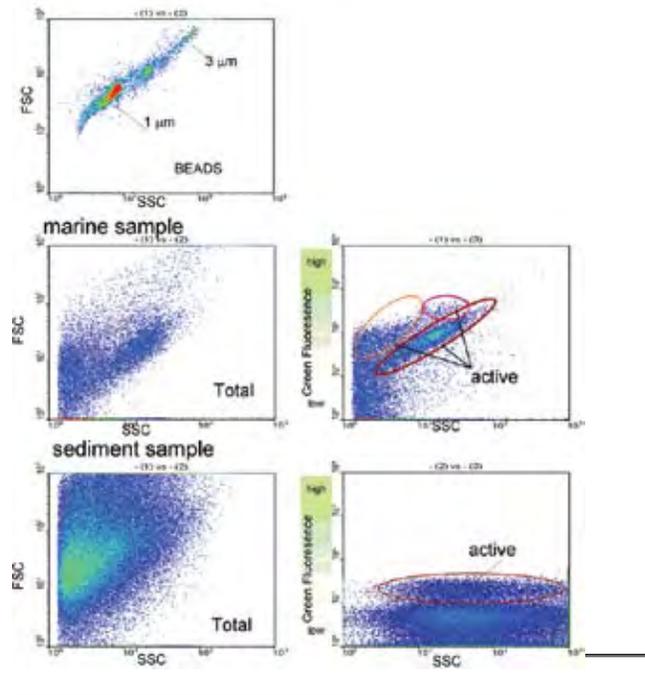


Figure 2. Flow cytometric analysis of cell populations from marine and sediment samples and stained with RSG.

We optimized the microwell design for bacteria and were easily able to detect O_2 consumption from individual microbial cells. This system was also successfully tested to measure the respiration rate of uncultured, live cells from the environment. We established a method for cell extraction and the examples of FC-plots of sediment and marine water samples are presented on Fig. 2. Microbial populations were separated by a CyFlowSpace flowcytometer/cell sorter (Partec); cells were then transferred onto the microchip platform and characterized in terms of respiration (O_2 consumption).

We demonstrated that these individual cells could be transferred to the 96 well plates using the Quixell Transfer micromanipulation system (Stoelting) and then used for whole-genome amplification (WGA) from single cells, and subsequent PCR analysis.

GTL

Proteomics Driven Analysis of Microbes, Plants and Microbial Communities

Mary S. Lipton^{1*} (mary.lipton@pnl.gov), Stephen J. Callister,¹ Joshua E. Turse,¹ Thomas Taverner,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 and Jim K. Fredrickson¹

¹Biological Sciences Division and ²Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Wash.

Collaborators: Lucy Shapiro (Stanford University); Derek Lovley (University of Massachusetts); Steve Giovannoni (Oregon State University); Sam Kaplan (UT-Houston Medical School); Jerry Tuscan (ORNL); and *Shewanella* Federation

<http://ober-proteomics.pnl.gov/>

Project Goals: The “High-Throughput Proteomic Analyses of Microbial and Plant Systems” 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 and plants.

Significance: Characterization of biological 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 higher organisms, 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 or utilizing plants as biofuels is the detailed understanding of the physiology of the cell. These cellular functions are dictated by the proteins expressed in the cell, their localization and their modification state. The “High-Throughput Proteomic Analyses of Microbial and Plant Systems” 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 and plants. In our poster, we highlight the ability to use proteomics data for genome annotation of microbes and fungi, characterization of microbial communities, advances in the characterization of protein phosphorylation state, and the identification of new proteins important to photosynthesis, and the determination of protein localization in stem, root and leaf tissues of poplar.

The proteome can play an integral role in the identification of protein-encoding genes (CDS) in sequenced genomes as part of the initial annotation of a genome. For example,

proteomics can be used to validate questionable gene identifications (e.g. encoding novel genes), identify CDS missed by automated gene calling algorithms, identify erroneous gene termini predictions, provide evidence for programmed frameshifting events that lead to alternative protein products, and provide evidence that sequencing mistakes are present. Up to now we have used proteomics data to enhance genome annotations after the initial annotation has been completed and publically released. In collaboration with JGI we are currently developing approaches to integrate proteomics data into the initial annotation pipeline. As part of this effort we have characterized numerous microbes and fungi that are targets of the GEBA genome sequencing project at the JGI to illustrate the utility of proteomic data in enhancing the quality of the annotation of genomic sequences and to allow subsequent cross-species comparative proteome analyses.

While comparative bacterial genomic studies commonly predict a set of genes indicative of common ancestry, experimental validation of the existence of this core genome requires extensive measurement and is typically not undertaken. Enabled by an extensive proteome database developed over six years, we have experimentally verified the expression of proteins predicted from genomic ortholog comparisons among 17 environmental and pathogenic bacteria. More exclusive relationships were observed among the expressed protein content of phenotypically related bacteria, which is indicative of the specific lifestyles associated with these organisms.

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

The *Rhodobacter sphaeroides* intracytoplasmic membrane (ICM) is an inducible membrane that is dedicated to the major events of bacterial photosynthesis, including harvesting light energy, separating primary charges, and transporting electrons. In this study, multichromatographic methods coupled with Fourier transform ion cyclotron resonance mass spectrometry and combined with subcellular fractionation, was used to test and prove the hypothesis that the photosynthetic membrane of *R. sphaeroides* 2.4.1 contains a significant number of heretofore unidentified proteins which are in addition to the integral membrane pigment-protein complexes previously discovered. These include light-harvesting complexes 1 and 2, the photochemical reaction center, and the cytochrome bc1 complex.

Our proteomic capabilities have been applied to characterize both the open ocean community in relation to *Pelagibacter 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* that a significant expression of the proteins involved in transport of metabolites and metals are indicative of the environmental metabolic requirements of this organism, and that this expression pattern is also represented in the larger sweater community.

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 courses 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, though 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/>

*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.

High Throughput Comprehensive Quantitative Proteomics and Metabolomics for Genomics:GTL

Gordon A. Anderson* (gordon@pnl.gov) Ronald J. Moore, David J. Anderson, Kenneth J. Auberry, Mikhail E. Belov, Stephen J. Callister, Therese R.W. Clauss, Kim K. Hixson, Gary R. Kiebel, Mary S. Lipton, Thomas O. Metz, Matthew E. Monroe, 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, Wash.

Project Goals: 1) Continue to operate and expand high throughput proteomics measurement capabilities;

2) Increase overall data quality, proteome coverage, and quantitative measurement precision/accuracy, and provide data with statistically sound measures of quality; 3) Characterize and implement technology advances that specifically augment Aims 1 and 2; and 4) Develop and improve capabilities for managing, disseminating, and mining proteomics results in support of Genomics:GTL-supported projects.

New and expanded capabilities for quantitative high throughput proteomics and metabolomics are being developed and applied to achieve high levels of data quality that enable broad studies, e.g., of diverse microbial systems, communities, and potentially ecosystems. In concert with other measurements and information, these capabilities are increasingly providing the basis for new systems-level biological insights.

Advancing the understanding of microbial and bioenergy-related systems is at the heart of DOE's Genomics:GTL program. A key aspect for acquiring such biological understandings at a systems level is the ability to quantitatively measure the array of proteins (i.e., the proteome) for a particular system under many different conditions, as well as the abundances of a large (and often uncertain) range of metabolites and other cellular components. Among the basic challenges associated with making useful comprehensive proteomic measurements are: 1) identifying and quantifying large sets of proteins whose relative abundances typically span many orders of magnitude, and 2) making proteomics measurements that provide information on protein-protein interactions and protein subcellular localization with sufficiently high throughput to enable practical systems biology approaches. Similar challenges exist for identifying and quantifying large sets of metabolites, in addition to even greater challenges for high throughput structural identification of metabolites due to broad structural diversity.

GTL

PNNL is addressing these issues by developing and applying new mass spectrometry-based measurement technologies in a high throughput environment to a range of collaborative GTL biological studies. Currently, our high throughput proteomics pipeline uses high resolution separations combined with mass spectrometry measurement capabilities that are integrated with advanced informatics tools amenable to the high data production rates and computational challenges associated with large-scale data comparisons. The accurate mass and time (AMT) tag strategy allows for both effective quantitative and high throughput peptide and intact protein-level analyses. Technological advancements related to sample processing and fractionation, as well as to the measurement platform itself have enabled an increasingly broad range of complex biological systems to be effectively addressed. Using the same mass spectrometry-based measurement platform, but with additional and complementary sample processing, fractionation and separation approaches, PNNL is also implementing complementary metabolomics approaches that aim to quantitatively define the broad range of other cellular constituents, and that are vital for understanding the function of biological systems.

Among the technological advancements for proteomics are new approaches that combine top-down and bottom-up measurements to extend quantitative proteome coverage to a large range of protein modification states. In another effort, a new fast separation LC-ion mobility-MS platform has been demonstrated to achieve high levels of data quality in conjunction with an order of magnitude increase in measurement throughput (see poster by R. D. Smith, et al.). These advances and new platform also directly benefit metabolomics, where measurements are complemented by the use of alternative separations and ionization modes in order to provide broad coverage of the diverse chemical universe of biological systems.

As part of these efforts, the supporting computational infrastructure at PNNL has also been expanded to incorporate a suite of search tools, data consolidation applications, and statistical relevance calculators, as well as visualization software for data interpretation. These informatics advances have enabled proteomic characterization of microbial communities from contaminated ground water and sediments, *Shewanella* strain comparisons, investigations of complex microbial response to oceanic seasonal cycling events, and more effective protein identifications, such as different modification states for *Caulobacter crescentis*. Moreover, these capabilities are also advancing GTL systems biology research; for example, proteomics measurements recently led to the discovery of active transport mechanisms involved in the uptake of nutrients in the oceans – a process previously thought to be governed by diffusion.

This poster will illustrate these technical advances, using specific applications to a range of Genomics:GTL projects as examples.

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.

—
GTL

A New Platform for Much Higher Throughput Comprehensive Quantitative Proteomics

Richard D. Smith* (rds@pnl.gov), Gordon A. Anderson, Erin Baker, Mikhail E. Belov, William F. Danielson III, Yehia M. Ibrahim, Ryan Kelly, Andrei V. Liyu, Matthew E. Monroe, Daniel J. Orton, Jason Page, David C. Prior, Keqi Tang, and Aleksey V. Tolmachev

Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Wash.

Project Goals: 1) Continue to operate and expand high throughput proteomics measurement capabilities; 2) Increase overall data quality, proteome coverage, and quantitative measurement precision/accuracy, and provide data with statistically sound measures of quality;

3) Characterize and implement technology advances that specifically; augment Aims 1 and 2; and 4) Develop and improve capabilities for managing, disseminating, and mining proteomics results in support of Genomics:GTL-supported projects.

Significance: A new fast separation LC-ion mobility-MS platform for quantitative high throughput proteomics measurements has been developed that can achieve high levels of data quality in conjunction with an order of magnitude increase in measurement throughput. The new platform not only provides improved sensitivity and significantly speeds large-scale applications, but also lowers the cost of proteomics measurements. As a result, previously impractical studies of diverse microbial systems, communities, and ecosystems, among others are now possible.

Obtaining a systems level understanding of complex systems such as microbial communities, and ecosystems requires characterization of very large numbers of samples (e.g., involving many perturbations or spatially and/or temporally distinct samples). Although the proteomics measurement throughput attainable with LC-MS(/MS)-based approaches is much greater than with classical (e.g., 2D-PAGE-MS) approaches, it generally is grossly short of that needed for many systems biology applications. To address this shortcoming, we have been developing a new platform with greatly improved measurement throughput, sensitivity, robustness, and quantitative capability for proteomics measurements applicable to a range of GTL program interests.

Our proteomics platform encompasses fast capillary LC separations coupled via a greatly improved electrospray ionization (ESI) interface to an ion mobility spectrometer (IMS) that is interfaced to a high speed and broad dynamic range time-of-flight mass spectrometer (TOF MS). A fully automated capillary LC system incorporates high pressure LC pumps, an autosampler, and a 4-column nanoscale fluidics system. Each 10-cm-long nanocapillary LC column is operated at 10,000 psi. ESI-generated ions are accumulated in an electrodynamic ion funnel trap before being injected into an 84-cm-long IMS drift tube. To increase IMS-TOF MS sensitivity, we developed a novel multiplexing approach that increases the number of ion pulses that can be separated in a given time by >50-fold. Downstream of the IMS drift tube, diffusion dispersed ion packets are collimated by an electrodynamic ion funnel into a high performance orthogonal acceleration TOF MS. A high-performance data acquisition was developed that enables high mass accuracy high dynamic range measurements.

Initial evaluation of the LC-IMS-TOF MS system platform has shown significantly improved performance compared to the best existing proteomics platforms. In our initial evaluation, encoded MS signals were reconstructed, “deisotoped,” and matched, using accurate mass and retention time information, against a reference database of peptides. The new platform was able to detect trace level peptides at high signal-to-noise ratios and an average peptide intensity coefficient of variance of ~9%. These results represent a significant improvement in data quality that is obtained in conjunction with the improvements in throughput.

Further improvements in performance from the use of an advanced ESI multiemitter source and ion funnel interface are expected to significantly build upon these initial results. In combination with improved informatics tools, this new proteomics capability is expected to enable previously impractical systems biology proteomics applications. Initial applications of the new platform are planned for the coming year.

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.

—
GTL

Live-Cell Visualization of Tagged Bacterial Protein Dynamics and Turnover

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

Pacific Northwest National Laboratory, Richland, Wash.

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. Accordingly, 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.**

Biarsenical multiuse affinity probes (MAPs) provide an important tool to determine the abundance, location, size, binding interfaces, and function of tagged protein in living cells (1). Complementary measurements using MAPs immobilized on solid supports permit the isolation of intact protein complexes replete with low-affinity binding partners for complementary functional and structural measurements (2, 3). Our recent introduction of multiple tagging sequences and complementary MAPs with different colors and improved photostability (4-6), coupled with the ability to make switchable MAPs for single-molecule measurements that permit subdiffraction imaging (7, 8), offers a

robust toolkit for high-throughput cellular measurements. However, while MAPs work well for the live-cell imaging of eukaryotic cells, their application in prokaryotes has been largely limited to the detection of highly overexpressed proteins in *E. coli* (9). Further, common protocols do not permit the stoichiometric labeling of tagged proteins, limiting their quantitative applications. To solve this latter hurdle in the application of MAPs in prokaryotes, we have systematically investigated methods that permit the specific labeling of tagged cytosolic proteins at near equimolar stoichiometries with minimal background signal from nonspecific labeling. Critical to our successful strategy was the expression of a tagged cyanofluorescent protein (CFP*) containing a C-terminus tag (i.e., CCPGCC), whose cytosolic location requires the delivery of the MAP (in this case FIAsh-EDT2) across the outer and inner membrane (Figure 1). Under optimal conditions, the tagged CFP* is selectively labeled; the small background signal for the empty-vector control is due to nonspecific labeling of endogenous bacterial proteins. Importantly, following expression of CFP*, which accounts for about 1% of cellular protein, nonspecific labeling is reduced. These results indicate that the tetra-cysteine tagging sequence on CFP successfully competes with nonspecific sites, resulting in specific labeling.



Figure 1. Live-Cell Visualization of Cytosolic Proteins. SDS-PAGE (left), live-cell images (center), and time-dependent labeling (right) showing *Shewanella oneidensis* MR-1 proteins following *in-vivo* labeling by FIAsh-EDT₂ (0.5 µM) for cells expressing an empty vector control (lanes 1 and 3; left panel or Control in center panel) or a CCPGCC tagged cyanofluorescent protein (CFP*) (lanes 2 and 4; left panel or Tagged Protein in center panel). SDS-PAGE (left) shows total protein stain (lanes 1 and 2) or fluorescence (lanes 3 and 4) for 40 µg total protein. Fluorescent images involve the direct excitation of FIAsh at 490 nm. Labeling Protocol. *Shewanella oneidensis* was cultured in LB medium at 30 °C; when the optical density reached 0.8, we simultaneously induced the expression of CFP* through the addition of arabinose (1 mM) and exposed cells to FIAsh (0.5 µM) in the presence of protamine (0.2 mg/mL) to enhance FIAsh uptake, and Disperse Blue 3 (20 µM) to reduce binding to nonspecific hydrophobic pockets. Labeling time is 2 hr. Cell Disruption. Cells were collected by centrifugation (5,000 x g), resuspended in PBS (50 mM NaH₂PO₄, 150 mM NaCl, pH 8.0), and cells were disrupted following the addition of lysozyme (5 mg/mL).

An example of the utilization of MAPs to access protein trafficking and turnover involved the metal reducing protein MtrC at the cell surface. These measurements took advantage of the probe CrAsH, whose net charge permits the selective labeling of tagged proteins located in the outer membrane. We find that MtrC is selectively labeled, and that trafficking to the surface requires the presence of the type II secretory system involving *gspG/gspD* (Figure 2). Expressed MtrC is stable under suboxic conditions; however, upon shifting to aerobic culture conditions the MtrC

proteins are degraded ($k_{\text{turnover}} = 0.036 \text{ hr}^{-1}$). The preferential degradation of MtrC under aerobic conditions is consistent with the sensitivity of this decaheme protein to oxidative damage, and suggests the presence of a protein degradative system that recognizes damaged proteins located on the extracellular face of the outer membrane.

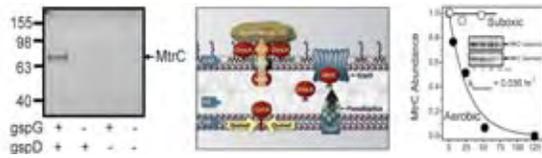


Figure 2. Specific Labeling of MtrC on cellular surface resolved using charged probe CrAsH (left panel), demonstrating that trafficking of MtrC involves type II secretion system (gspG and gspD) (center panel). Half-life of CrAsH-labeled MtrC is dependent on environmental conditions (right panel), consistent with role of MtrC in mediating electron transfer under suboxic conditions.

Conclusions:

- Equimolar stoichiometric labeling of tagged cytosolic proteins permits quantitative live cell measurements of bacterial protein dynamics.
- Selective labeling of outer membrane proteins permits measurements of trafficking and structures of membrane proteins.

References

1. Adams, S. R., Campbell, R. E., Gross, L. A., Martin, B. R., Walkup, G. K., Yao, Y., Llopis, J. and Tsien, R. Y. (2002) New biarsenical ligands and tetracysteine motifs for protein labeling in vitro and in vivo: synthesis and biological applications. *Journal of the American Chemical Society* 124, 6063-6076.
2. Mayer, M. U., Shi, L. and Squier, T. C. (2005) One-step, non-denaturing isolation of an RNA polymerase enzyme complex using an improved multi-use affinity probe resin. *Molecular Biosystems* 1, 53-56.
3. Verma, S., Xiong, Y., Mayer, M. U. and Squier, T. C. (2007) Remodeling of the bacterial RNA polymerase supramolecular complex in response to environmental conditions. *Biochemistry* 46, 3023-3035.
4. Cao, H., Xiong, Y., Wang, T., Chen, B., Squier, T. C. and Mayer, M. U. (2007) A red cy3-based biarsenical fluorescent probe targeted to a complementary binding peptide. *Journal of the American Chemical Society* 129, 8672-8673.
5. Chen, B., Cao, H., Yan, P., Mayer, M. U. and Squier, T. C. (2007) Identification of an orthogonal peptide binding motif for biarsenical multiuse affinity probes. *Bioconjugate chemistry* 18, 1259-1265.
6. Wang, T., Yan, P., Squier, T. C. and Mayer, M. U. (2007) Prospecting the proteome: identification of naturally occurring binding motifs for biarsenical probes. *ChemBiochem* 8, 1937-1940.
7. Bates, M., Huang, B., Dempsey, G. T. and Zhuang, X. (2007) Multicolor Super-Resolution Imaging with Photo-Switchable Fluorescent Probes. *Science*.
8. Rust, M. J., Bates, M. and Zhuang, X. (2006) Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature methods* 3, 793-795.

9. Chen, B., Mayer, M. U., Markillie, L. M., Stenoien, D. L. and Squier, T. C. (2005) Dynamic motion of helix A in the amino-terminal domain of calmodulin is stabilized upon calcium activation. *Biochemistry* 44, 905-914.

GTL

Identification of All Engineering Interventions Leading to Targeted Overproductions

Sridhar Ranganathan* (sur152@psu.edu), Patrick F. Suthers, and **Costas D. Maranas** (costas@psu.edu)

Dept. of Chemical Engineering, Pennsylvania State University, University Park, Pa.

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

Project Goals: The general aims of this project are: (Aim 1) to generate integrated computational tools for the automated generation and curation of genome-scale models of metabolism for microbial and plant systems; (Aim 2) to automatically generate maps tracking the fate of labeled isotopes through genome-scale models; (Aim 3) to fully elucidate metabolic fluxes in genome-scale models using GC-MS or NMR data; (Aim 4) to leverage flux data for wild-type strain to identify all possible engineering strategies that lead to overproduction of a targeted product.

Existing computational methods (e.g., OptKnock, METAOPT, OptGene) for strain redesign generate engineering interventions only one at a time, limiting the array of choices presented to the biotechnologist. Metabolic flux data obtained through high-throughput experiments (e.g., MFA) are not used directly, instead the maximization of a surrogate of cellular fitness (i.e., max of biomass) is employed to generate metabolic flux predictions. To remedy these limitations, we present a computational framework that predicts all possible engineering strategies for the overproduction of targeted biochemicals. Instead of looking at engineering strategies one at a time, we identify engineering interventions by classifying reactions in the metabolic model depending upon whether their flux values must increase, decrease or become equal to zero to meet the pre-specified overproduction target. We hierarchically apply this classification rule for pairs, triples, quadruples, etc. of reactions. This leads to the identification of a sufficient and non-redundant set of fluxes that must change (i.e., MUST set) to meet a pre-specified overproduction target. Starting with this MUST we subsequently extract a minimal subset of fluxes that must actively be engineered (i.e., FORCE set) to ensure that all fluxes in the network are consistent with the overproduction objective.

We first demonstrated the developed methodology for succinate production in *Escherichia coli* using the most recent *in silico* *E. coli* model, iAF1260. The method not only recapitulates existing engineering strategies (e.g., overexpression of phosphoenolpyruvate carboxylase and elimination of the competing byproduct ethanol) but also reveals non-intuitive

ones that boost succinate production by performing coordinated changes on pathways distant from the last steps of succinate synthesis. Following this study, we addressed the overproduction of various biochemicals identified as promising biofuels.

GTL

Automated Construction of Genome-Scale Metabolic Models: Application to *Mycoplasma genitalium*

Patrick F. Suthers^{1*} (suthers@engr.psu.edu), Vinay Satish Kumar,² and **Costas D. Maranas**¹ (costas@psu.edu)

¹Dept. of Chemical Engineering and ²Dept. of Industrial Engineering, Pennsylvania State University, University Park, Pa.

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

Project Goals: The general aims of this project are: (Aim 1) to generate integrated computational tools for the automated generation and curation of genome-scale models of metabolism for microbial and plant systems; (Aim 2) to automatically generate maps tracking the fate of labeled isotopes through genome-scale models; (Aim 3) to fully elucidate metabolic fluxes in genome-scale models using GC-MS or NMR data; (Aim 4) to leverage flux data for wild-type strain to identify all possible engineering strategies that lead to overproduction of a targeted product.

Currently, over 700 genomes (including eleven plant species) have been fully sequenced, however, only about 20 organism-specific genome-scale metabolic models have been constructed. It appears that metabolic model generation can only keep pace with about 1% of the fully sequenced genomes. In response to this flood of present and future genomic information, automated tools such as Pathway Tools and SimPhenyTM have been developed that, using homology comparisons, allow for the automated generation of draft organism-specific metabolic models. This has shifted the burden towards curating the accuracy and completeness of the automatically generated, though draft, reconstructions. All of these reconstructions remain to some extent incomplete as manifested by the presence of unreachable metabolites and growth inconsistencies between model predictions and observed *in vivo* behavior.

In this work, we highlight the construction of *in silico* metabolic models for a minimal organism *Mycoplasma genitalium* [1]. Key challenges for *M. genitalium* included estimation of biomass composition, handling of enzymes with broad specificities and the lack of a defined medium. Our computational tools were employed to identify and resolve connectivity gaps in the model (*i.e.*, GapFind and GapFill) as well as growth prediction inconsistencies with gene essentiality experimental data (*i.e.*, GrowMatch). The resulting curated model, *M. genitalium* iPS189 (262 reactions, 274 metabolites) was 87% accurate in recapitulating *in vivo* gene

essentiality results for *M. genitalium*. Specifically, we found that the model correctly identified 149 out of a total of 171 essential genes (*i.e.*, specificity of 87%) and 16 out of a total of 18 non-essential genes (*i.e.*, sensitivity of 89%). This level of agreement not only meets, but exceeds thresholds for metabolic model quality put forth in the literature. Additional *in vivo* gene essentiality studies using a fully defined medium could usher a more accurate elucidation of the true metabolic capabilities of *M. genitalium*, as well as suggest improvements to the model.

We describe the application of iPS189 to drive the development of a defined growth medium. As noted above, gene essentiality experiments were performed using a non-defined medium, SP-4, which contains beef heart infusion, peptone supplemented with yeast extract and fetal bovine serum. The use of undefined medium can confound the characterization of gene essentiality, as the exact environmental conditions are not fully specified. Motivated by these shortcomings, we used the iPS189 metabolic model as a roadmap of the available transporters, metabolites and internal interconversions to seek out growth medium components necessary for biomass production.

Reference

1. Suthers, PF, MS Dasika, V Satish Kumar, G Denisov, JI Glass and CD Maranas (2008), *PLoS Comp Biol*, accepted

GTL

Synthetic Lethality Analysis Based on the *Escherichia coli* Metabolic Model

Alireza Zomorodi* (auz107@psu.edu), Patrick F. Suthers, and **Costas D. Maranas** (costas@psu.edu)

Dept. of Chemical Engineering, Pennsylvania State University, University Park, Pa.

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

Project Goals: The general aims of this project are: (Aim 1) to generate integrated computational tools for the automated generation and curation of genome-scale models of metabolism for microbial and plant systems; (Aim 2) to automatically generate maps tracking the fate of labeled isotopes through genome-scale models; (Aim 3) to fully elucidate metabolic fluxes in genome-scale models using GC-MS or NMR data; (Aim 4) to leverage flux data for wild-type strain to identify all possible engineering strategies that lead to overproduction of a targeted product.

Essential genes are defined as genes whose deletion is lethal. By analogy, synthetic lethals (SL) refer to pairs of non-essential genes whose simultaneous deletion negates biomass formation. One can extend the concept of lethality by considering gene groups of increasing size where only the simultaneous elimination of all genes is lethal whereas individual gene deletions are not. Synthetic lethality results provide a birds-eye view of the redundant mechanisms avail-

able for redirecting metabolism and reveal complex patterns of gene utilization and interdependence. We developed a bilevel optimization-based procedure for the targeted enumeration of all multi-gene (and by extension multi-reaction) lethals for genome-scale metabolic models that outperforms exhaustive enumeration schemes by many orders of magnitude.

This proposed synthetic lethality analysis is applied to the *iAF1260* model of *E. coli* K12 for aerobic growth on minimal glucose medium leading to the identification of all single, double, triple, quadruple and some higher-order SLs. The identified gene/reaction synthetic lethal pairs are phenotypically classified into two types: those yielding auxotroph strains that can be rescued through the supply of missing nutrients (i.e., amino acids or other compounds), and the ones lacking essential functionalities that cannot be restored by adding extra components to the growth medium. Graph representations of these synthetic lethals reveal a variety of motifs ranging from disjoint pairs, to hub-like stars, to *k*-connected sub-graphs. By analyzing the functional classifications of the genes involved in synthetic lethals (such as carbohydrate or nucleotide metabolism)

we uncover trends in connectivity within and across COG functional classifications. We contrast the identified synthetic lethal predictions against experimental data and suggest a number of model refinement possibilities as a direct consequence of the obtained results.

We exhaustively assessed SL reaction triples and identified all reactions participating in at least one SL quadruples. The concept of degree of essentiality is introduced, defined as the lowest degree of a SL that a reaction participates, to unravel the contribution of each reaction in “buffering” cellular functionalities. We find that reactions in different COG classifications often involve very different degree of essentiality statistics. This study provides a complete analysis of gene essentiality and lethality for the latest *E. coli iAF1260* and ushers the computational means for performing similar analyses for other genome-scale models. Furthermore, by exhaustively elucidating all model growth predictions in response to multiple gene knock-outs it provides a many-fold increase in the number of genetic perturbations that can be used to assess the performance of *in silico* metabolic models.

Molecular Interactions and Protein Complexes

The MAGGIE Project

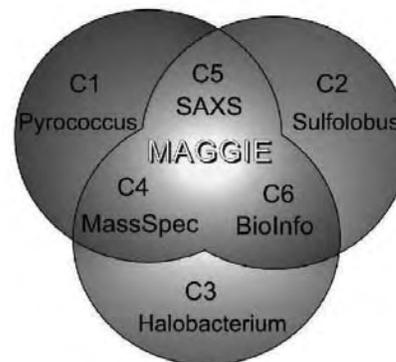
High Throughput Structural Characterization of Protein Complexes in Solution using Small Angle X-ray Scattering (SAXS) Combined with Mass Spectrometry (MS)

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. Baliga⁴

¹Life Science & Physical Biosciences Divisions, Lawrence Berkeley National Laboratory, Berkeley, Calif.; ²University of Georgia, Athens, Ga.; ³The Scripps Research Institute, La Jolla, Calif.; and ⁴Institute for Systems Biology, Seattle, Wash.

Project Goals: This is a Genomics:GTL proposal to support solution x-ray structures and other technologies to characterize Protein Complexes and Modified Proteins in microbes.

—
GTL



The GTL Project MAGGIE (Molecular Assemblies, Genes, and Genomics Integrated Efficiently) integrates teams at Lawrence Berkeley National Lab and the Advanced Light Source (ALS) with other researchers to characterize the Protein Complexes (PCs) and Modified Proteins (MPs) controlling microbial cell biology including stress and metabolic pathways relevant to bio-energy. Two key tools are Small Angle X-ray Scattering (SAXS) and Mass Spectrometry (MS). SAXS informs folding, unfolding, aggregation, shape, conformation, and assembly state in solution. SAXS resolution is limited to 50 to 10 Å resolution, but escapes the size limitations inherent in NMR and electron microscopy studies. We designed and built the SIBYLS synchrotron beamline (<http://www.bl1231.als.lbl.gov>) at the Advanced Light Source to interconvert between a SAXS and a crystallography endstation quickly allowing combined techniques. SAXS experiments and

results on native microbial biomass combined with MX and MS reveals the power of these combined techniques. SAXS results can determine correct molecular mechanisms for complexes involving conformational changes (Yamagata and Tainer, 2007). SAXS can identify extended or unstructured regions, and provide information on every sample at relatively high throughput (Putnam et al., 2007). The methods discussed provide the basis to examine molecular complexes and conformational changes relevant for accurate understanding, simulation, and prediction of mechanisms in structural cell biology and nanotechnology.

References

1. Putnam, C.D., Hammel, M., Hura, G. L., and Tainer, J. A. (2007) X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution, *Quarterly Reviews in Biophysics* 40,191-285.
2. Yamagata, A, and Tainer, JA (2007). "Hexameric structures of the archaeal secretion ATPase GspE and implications for a universal secretion mechanism." *The EMBO Journal* 26, 878-890.

GTL

Define the Metalloproteome by Letting Metals Take the Lead: A Component of the MAGGIE Project

Aleksandar Cvetkovic^{1*} (alcv@uga.edu), Angeli Lal Menon,¹ Farris L. Poole II,¹ Sarat Shanmukh,¹ Michael Thorgersen,¹ Joseph Scott,¹ Jeremy Praissman,¹ Ewa Kalisiak,² Sunia Trauger,² Steven M. Yannone,³ John A. Tainer,³ Gary Siuzdak,² and **Michael W.W. Adams¹**

¹Dept. of Biochemistry and Molecular Biology, University of Georgia, Athens, Ga.; and ²Center for Mass Spectrometry and ³Dept. of Biochemistry and Molecular Biology, The Scripps Research Institute, La Jolla, Calif.

Project Goals: The overall goal of the MAGGIE project (Molecular Assemblies, Genes and Genomes Integrated Efficiently) is to provide robust GTL technologies and comprehensive characterization to efficiently couple gene sequence and genomic analyses with protein interactions and thereby elucidate functional relationships and pathways. The operational principle guiding MAGGIE objectives is that protein functional relationships involve interaction mosaics that self-assemble from independent protein pieces that are tuned by modifications and metabolites, including metals. The objective is therefore to comprehensively characterize the Protein Complexes (PCs) and Modified Proteins (MPs), including metalloproteins, underlying microbial cell biology.

Metal ion co-factors afford proteins with virtually unlimited catalytic potential, enable electron transfer reactions and are major determinants of protein stability. Consequently, metalloproteins play key roles in most if not all biological processes. However, it is not possible to predict the types of

metal that an organism uses or the number and/or types of MPs encoded in its genome sequence because metal coordination sites are diverse and not easily recognized. The metalloproteome of an organism has yet to be defined. Directly determining the identity of MPs in native biomass utilizing a novel approach which focuses on the identification and purification of a metal instead of a protein which addresses some of these issues is described. 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 provide both stable and dynamic protein complexes and metalloproteins for further characterization.

Liquid chromatography and high-throughput tandem mass spectrometry were used to separate and identify proteins, and metals were identified by inductively coupled plasma mass spectrometry (ICP-MS). After fractionation of a cytoplasmic extract from *P. furiosus*, a total of 249 metal peaks were identified, 154 of which appear to be unknown metalloproteins. These uncharacterized peaks included metals (uranium, lead, molybdenum, manganese and vanadium) that the organism was not previously known to utilize, as well as metals that were (iron, nickel, cobalt, tungsten and zinc). Similar chromatographic and metal analyses were performed for detergent-solubilized membranes of *P. furiosus* and for the cytoplasm of two other microorganisms, *Sulfolobus solfataricus* and *Escherichia coli* (grown on their conventional laboratory media). These revealed peaks of yet other types of metal, including tin, cadmium, zirconium, arsenic and thallium, all associated with as yet unknown proteins. Identification of these novel metalloproteins is underway and is being facilitated by the development of HPLC analyses with in-line ICP-MS. Our overall goal is to develop an analytical approach to define the metalloproteome of any organism under any growth condition.

These results indicate that metalloproteomes are much more extensive than previously recognized, and likely involve both biologically conventional and unanticipated metals with implications for a complete understanding of cell biology.

This research was funded by the Department of Energy (DE-FG0207ER64326) as part of the MAGGIE project.

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

Robert Rambo,¹ Trent Northen,¹ Adam Barnebay,¹ Kenneth Stedman,² Michael W.W. Adams,³ Gary Siuzdak,⁴ Nitin S. Baliga,⁵ Steven R. Holbrook,¹ John A. Tainer,^{1,6} and **Steven M. Yannone**^{1*} (SMYannone@lbl.gov)

¹Dept. of Molecular Biology, Lawrence Berkeley National Laboratory, Berkeley, Calif.; ²Center for Life in Extreme Environments, Portland State University, Portland, Wash.; ³Dept. of Biochemistry and Molecular Biology, University of Georgia, Athens, Ga.; ⁴Center for Mass Spectrometry, The Scripps Research Institute, La Jolla, Calif.; ⁵Institute for Systems Biology, Seattle, Wash.; and ⁶Dept. of Biochemistry and Molecular Biology, The Scripps Research Institute, La Jolla, Calif.

Project Goals: As component 2 of the 6 component MAGGIE project our goals include 1) providing highly purified molecular machines from *Sulfolobus solfataricus* for physical characterization by SAXS (Component 5) and identification by MS/MS (Component 4), 2) Developing molecular biology in *Sulfolobus solfataricus* for recombinant protein expression in *Sulfolobus*, 3) Develop non-disruptive biochemical techniques to isolate and identify intact molecular machines from native biomass. 4) Develop methods to isolate and identify membrane protein assemblies and complexes. Ultimately, we aim to identify metabolic modules suitable to transfer specific DOE missions while developing generally applicable molecular and biophysical technologies for GTL.

Dynamic protein-protein interactions are fundamental to most biological processes and essential for maintaining homeostasis within all living organisms. These interactions create dynamic and diverse functional networks essential to biological processes. Thus, a thorough understanding of these networks will be critical to engineering biological processes for DOE missions. The MAGGIE project was conceived, in part, as a response to the DOE GTL initiative to develop technologies to map the proteomes of model organisms. In this project we are exploiting unique characteristics of members of extremophilic Archaea to identify, isolate, and characterize multi-protein molecular machines. We have teamed expertise in mass spectrometry, systems biology, structural biology, biochemistry, and molecular biology to approach the challenges of mapping relatively simple proteomes. As part of the MAGGIE project, we are developing shuttle vectors for the extremophilic organism *Sulfolobus solfataricus* which has a growth optimum at 80°C and pH 3.0. We are using naturally occurring viral pathogens and

plasmids to engineer shuttle vectors designed for recombinant protein tagging and expression in the native *Sulfolobus* background. The MAGGIE project has developed broadly applicable technologies for cellular deconstruction, isolation, and identification of intact molecular machines from native biomass. Our approach maintains a native environment throughout the characterization of the protein complexes and includes the complete partitioning of membrane proteins from the “soluble” material. Stable complexes are further characterized by small angle X-ray scattering (SAXS) at the Advanced Light Source at LBNL which provides the overall shape and size of the complex in solution. This work is directed toward developing rapid and efficient means to identify, isolate, and characterize molecular machines from any organism by integrating biochemical techniques with cutting-edge mass spectrometry, X-ray scattering and bioinformatics in a single approach. 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.

Technologies for Robust and Rational Reengineering of Microbial Systems

W.L. Pang,^{1*} T. Koide,¹ D.J. Reiss,¹ C. Bare,¹ K. Beer,¹ R. Bonneau,² S. Coesel,¹ M.T. Facciotti,³ A. Kaur,¹ F.Y. Lo,¹ K. Masumura,⁴ D. Miller,¹ M. Orellana,¹ M. Pan,¹ A.K. Schmid,¹ P. Shannon,¹ D. Tenenbaum,¹ P.T. Van,¹ K. Whitehead,¹ and **N.S. Baliga**¹ (nbaliga@systemsbiology.org)

¹Institute for Systems Biology, Seattle, Wash.; ²New York University, New York, N.Y.; ³University of California, Davis, Calif.; and ⁴National Institute for Health Sciences, Tokyo, Japan

Project Goals: Component 3: Bolster through high-end state-of-art systems approaches, developed specifically for the study of archaeal organisms, the comprehensive analysis of multi-protein complexes in DOE-relevant organisms.

Technologies to synthesize and transplant a complete genome into a cell have opened limitless potential to redesign organisms for complex specialized tasks. However, large scale reengineering of a biological circuit will require systems level optimization that will come from a deep understanding of operational relationships among all of a cell's constituent parts. We have developed systems approaches for the global deconstruction of transcriptional networks into statistically predictive models. We have also developed microfluidic platforms for reductionist analysis of sub-circuits within this global model to generate high-resolution kinetic expression

profiles that aid the development of fine-grained ODE/SDE based predictive simulations. The coupling of these two otherwise diametrically opposed approaches will provide avenues for the type of multiscale modeling necessary for systems reengineering.

Dynamic Assembly of Functional Transcriptional Complexes Inside Genes and Operons

David J. Reiss* (dreiss@systemsbiology.org), Tie Koide, W. Lee Pang, J. Christopher Bare, Marc T. Facciotti, Amy K. Schmid, Min Pan, Bruz Marzolf, Phu T. Van, Fang-Yin Lo, Abhishek Pratap, Eric W. Deutsch, Amelia Peterson, Dan Martin, and **Nitin S. Baliga** (nbaliga@systemsbiology.org)

Institute for Systems Biology, Seattle, Wash.

Project Goals: Component 3: Bolster through high-end state-of-art systems approaches, developed specifically for the study of archaeal organisms, the comprehensive analysis of multi-protein complexes in DOE-relevant organisms.

Transcription complexes assemble dynamically at specific sites on DNA to drive the controlled transcription of downstream genes. Despite knowledge of complex prokaryotic transcription mechanisms, generalized rules, such as the simplified organization of genes into operons with well-defined promoters and terminators, have played a significant role in systems analysis of regulatory logic in both bacteria and archaea. Through integrated analysis of transcriptome dynamics and protein-DNA interactions measured at high resolution throughout the *Halobacterium salinarum* NRC-1 genome, we have identified widespread environment-dependent modulation of operon architectures, transcription initiation and termination inside coding sequences, and extensive overlap in 3' ends of transcripts for many convergently transcribed genes. We demonstrate that a significant fraction of these alternate transcriptional events correlate to binding locations of 11 transcription factors (TFs) inside operons and annotated genes – events often considered spurious or non-functional. This has illustrated the prevalence of overlapping genomic signals in archaeal transcription.

Protein Complex Analysis Project (PCAP)

GTL

GTL

Protein Complex Analysis Project (PCAP): Automated Particle Picking in Electron Microscopy Images

Pablo Arbelaez^{1*} (arbelaez@eecs.berkeley.edu), Bong-Gyoon Han,^{2*} (bghan@lbl.gov), Robert M. Glaeser,² and Jitendra Malik¹

¹University of California, Berkeley, Calif.; and ²Lawrence Berkeley National Laboratory, Berkeley, Calif.

Project Goals: Develop a computer system capable of detecting automatically protein complexes in Electron Microscopy images. The method should be robust enough to process a large variety of particles and the quality of results should be comparable to human picking. Performance will be measured by the resolution of the final three dimensional reconstruction, as the main goal of the project is to propose a building block towards a fully automatic pipeline in Single Particle Analysis.

Particle picking is one of the main bottlenecks towards automating Single Particle Analysis. In this work, we propose a general framework for automatic detection of molecular structures in Electron Microscopy images. For this purpose, we formulate the problem as a visual pattern recognition task and use texture information as the main perceptual cue to differentiate particles from background. We follow a discriminative approach and train a non-linear classifier in order to precisely localize the particles in micrographs. Experiments conducted on several datasets of negatively stained specimens show that the reconstructions obtained from particles picked by our algorithm have resolution comparable with the ones obtained from human-picked particles.

GTL

A High Throughput Pipeline for Identifying Protein-Protein Interactions in *Desulfovibrio vulgaris* using Tandem-Affinity Purification

Swapnil Chhabra,¹ Gareth Butland^{1*} (GPButland@lbl.gov), Dwayne Elias,² Veronica Fok,¹ Ramadevi Prathapam,¹ Thomas Juba,² John Marc Chandonia,¹ Ewa Witkowska,³ **Mark Biggin**¹ (MDBiggin@lbl.gov), Terry Hazen,¹ Judy Wall,² and Jay Keasling^{1,4}

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

Project Goals: The broad goal of this subproject of PCAP is to study systems level biology of microbes of interest to DOE by developing a high throughput pipeline for facile genetic manipulation of these organisms. Specifically our goal is determine the genome-wide network of protein interactions, the abundance and spatial organization of the protein complexes within *Desulfovibrio vulgaris* Hildenborough (DvH), under normal laboratory and environmentally relevant stress conditions. These network inference data will be used to build comprehensive cellular models of metal reducing bacteria. The challenges faced in this project included the choice of a strictly anaerobic bacterium as the model soil bacterium, rudimentary genetic tools, and the effects of a high sulfide environment.

Most cellular processes are mediated by multiple proteins interacting with each other in the form of multi-protein complexes and not by individual proteins acting in isolation. In order to extend our functional genomics analyses of stress response pathways in *Desulfovibrio vulgaris*, we want to study the role of protein complexes in this sulfate reducing bacterium which has been found to exist in several DOE waste sites. Here we report the development of a technological platform for rapid identification of protein-protein interactions from a library of *D. vulgaris* mutant strains expressing tagged proteins. Our existing platform is based on the single cross-over chromosomal integration of tagged constructs generated in *E. coli* and we demonstrate the successful implementation of tagged-strain generation, verification and identification of interacting protein partners.

We demonstrate the strain generation process using automated software and hardware tools such as LIMS for automated sequence alignments and strain tracking, liquid handling systems for processing nucleic acids. We generated 363 tagged (STF/SPA) clones using the two-step TOPO-Gateway[®] cloning approach (Invitrogen) of which 278 were electroporated into competent *D. vulgaris* cells. We confirmed the single cross-over integration of 76 strains and the expression of affinity tagged fusion proteins using anti-FLAG IP-western blots and the verified strains were then subjected to TAP purification and MS based

identification of interacting proteins. The TOPO-Gateway[®] mediated single-cross-over approach works best for genes located at terminal ends of operons. In order to overcome these limitations, we have further developed the use of a double cross-over approach mediated through Sequence and Ligation Independent Cloning (SLIC). We will report our efforts towards adapting this method into a high throughput affinity tagged strain generation platform applicable to many species of interest to DOE.

GTL

Protein Complex Analysis Project (PCAP): Protein Complex Purification and Identification by “Tagless” Strategy

Ming Dong,^{1*} Haichuan Liu,² Lee Yang,¹ Megan Choi,¹ Evelin D. Szakal,² Simon Allen,² Steven C. Hall,² Susan J. Fisher,^{1,2} Gareth Butland,¹ Terry C. Hazen,¹ Jil T. Geller,¹ Mary E. Singer,¹ Peter Walian,¹ Bing Jap,¹ Jian Jin,¹ John-Marc Chandonia,¹ H. Ewa Witkowska,² and **Mark D. Biggin**¹ (mdbiggin@lbl.gov)

¹Lawrence Berkeley National Laboratory, Berkeley, Calif.; and ²University of California, San Francisco, Calif.

Project Goals: Water soluble protein complexes identified from 1% of the DvH proteome using the tagless strategy are presented. The data demonstrate that approximately 70% of polypeptides participate in multimeric assemblages. Multidimensional clustering methods are being developed to provide higher confidence for assigning putative heteromeric complexes based on similarity of polypeptide elution profiles. Many DvH homomers demonstrate unique stoichiometries that differ from those of their close orthologs.

Subproject B 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. One of them is a “tagless” strategy for purifying and identifying water soluble protein complexes. The tagless strategy employs multi-dimensional separation of protein complexes followed by mass spectrometric monitoring of their composition. First, a crude cell extract (400 L culture) is separated under native conditions into a large number of fractions by an orthogonal four-step purification scheme. Then a rational sampling of ~30,000 of these fractions is made for MS analysis. Proteins are denatured, digested with trypsin, and then derivatized with primary amine-specific iTRAQ reagents. Samples are multiplexed (4- or 8-plexed, depending on the iTRAQ reagent used) and analyzed by LC MALDI MS/MS. In the course of MS/MS analysis, polypeptides are identified and, at the same time, their relative abundances are measured. As a consequence of the iTRAQ-based quantitation, elution profiles are generated for each polypeptide as it migrates through the column. We employ a clustering algorithm that

selects subsets of polypeptides that exhibit similar elution profiles to identify candidate heteromeric complexes. Polypeptides eluting without obvious partners at significantly higher MW than predicted from genome sequence data are categorized as putative homomers.

Over the last year, we have introduced further improvements to the tagless protocol, for example a miniaturized version of the 96-well PVDF multi-screen plate-based iTRAQ labeling procedure that lowers reagent cost by roughly a factor of ten. In addition, we have also developed an iterative MS/MS acquisition (IMMA) to facilitate greater sample analysis throughput by accelerating the rate limiting MS/MS analysis step (see additional poster).

To date, MS analysis of this pipeline has been performed on 395 fractions from 20 hydrophobic interaction columns that were in turn derived from 2 neighboring Mono Q column fractions. This represents around 1% of the total *D. vulgaris* water soluble protein fractionation space. Among 467 proteins identified, 263 were identified on the basis of 3 or more peptides, 343 on the basis of 2 peptides and the remaining 124 on the basis of 1 peptide. As judged by their elution profiles, the majority of polypeptides (~70%) migrated through a size exclusion column as multimeric assemblages. Seventeen heteromeric complexes were inferred on the basis of their homology to *E. coli* orthologs or a shared operon. In addition, many other polypeptides were found to cluster. Thus, to increase the level of confidence regarding putative interactions of polypeptides, a multidimensional clustering analysis algorithm aimed at comparing polypeptide elution profiles in all neighboring fractions is currently being developed. A total of 176 putative homomeric complexes were identified. One hundred of these have orthologs in *E. coli*, 55 of which have stoichiometries that differ between the two species. Independent analysis by electron microscopy of a number of these *DvH* homomeric complexes confirms that their stoichiometries differ not only from those of *E. coli* complexes but also in some cases from those of orthologs in more closely related species.

—
GTL

Protein Complex Analysis Project (PCAP): Localization of Multi-Protein Complexes through SNAP-Tag Labeling

Manfred Auer,^{1,3*} **Mark D. Biggin**^{1,3} (MDBiggin@lbl.gov),
Gareth Butland,^{1,3} Swapnil Chhabra,^{1,3} Dwayne A. Elias^{2,3}
(eliasd@missouri.edu), Afolami Fagorala,³ Terry Hazen,^{1,3}
Danielle Jorgens,^{1,3} Dominique C. Joyner,³ Thomas R.
Juba,^{2,3} Melissa Perez,^{1,3} Jonathan P. Remis,^{1,3} Andrew
Tauscher,^{1,3} and Judy D. Wall^{2,3}

¹Lawrence Berkeley National Laboratory, Berkeley,
California; ²University of Missouri, Columbia, Mo.;
and ³Virtual Institute for Microbial Stress and Survival,
Berkeley, Calif.

<http://vimss.lbl.gov/projects/pcap.html>

Project Goals: The broad goal of this subproject of PCAP is to develop tag-based labeling approaches for high-throughput subcellular localization of proteins in microorganisms of interest to DOE. Our goal is to determine the abundance, the spatial organization and relative locations of proteins within individual, intact *Desulfovibrio vulgaris Hildenborough* (DvH) cells, as well as in microbial communities, under normal laboratory and environmentally relevant stress conditions. Further, we are attempting to correlate the intracellular abundance of proteins with toxic metal reduction activity. The challenges faced in this project included the choice of a strictly anaerobic bacterium as the model soil bacterium, rudimentary genetic tools, application of a fluorescent protein tag in an anaerobic bacterium for the first time and the effects of a high sulfide environment.

The key to an effective high-throughput approach has been the development of a functional genetic tagging approach. We chose to introduce tags onto single copies of the genes encoding the target proteins that were regulated by wild-type promoter sequences.

In initial attempts to visualize fluorescent tags, we encountered background labeling-problems with ReAsH labeling (Invitrogen) and, as such, have concentrated our efforts on SNAP-labeling (Covalys Technologies) because several SNAP-tagged *D. vulgaris* strains were available. Both the commercially available Gateway vectors and Sequence and Ligation Independent Cloning (SLiC) plasmids and methodologies were used to ligate the SNAP tag onto the carboxy terminal end of the gene of interest. Once the gene, tag and the direction of insertion into the plasmid were verified with PCR-based sequencing, the plasmid was electroporated into *D. vulgaris*. The resulting colonies were picked, checked for contamination, and the incorporation of the gene and tag into the *D. vulgaris* genome verified with PCR screening. If positive, cells putatively containing tagged constructs were grown and cell lysates subjected to an affinity/SNAP antibody column. The eluted protein was visualized on a denaturing gel and, if a protein band of the predicted molecular weight was detected with the SNAP antibody, this construct was considered confirmed. To date, it appears that the SLiC methodology for tagged gene plasmid production is cheaper, faster and has the advantage of ease of introduction of a promoter for expression of downstream genes in an operon.

The remaining challenges include a determination of the optimal screening and verification methods for the tagged construct. In some cases, Southern blots were performed and showed that the correct genetic structure had been achieved, yet no SNAP IP product could be detected. Conversely, cases existed where SNAP IP analysis was positive, but there was an absence of fluorescent labeling with either *in vivo* or *in vitro* labeling using SDS-PAGE analysis. In addition there has been heterogeneity of labeling among genetically identical bacteria, as well as our localization results in planktonic cells and biofilms using photoconversion approaches.

As a proof of principle for our fluorescent tag visualization, we have extended our approach to a small number of *E. coli*

constructs where the localization of the proteins are better established. These experiments will give us confidence that our procedures are working correctly. Once the labeling and photoconversion protocol are sufficiently robust, they can be applied to a growing number of DvH tagged strains under a variety of environmental conditions, including stress conditions. Once of the things we hope to learn is whether differences in protein expression levels, that are apparent in planktonic cells, may be the reason for cell-to-cell differences of metal reduction capability as seen in planktonic cells and in biofilms.

GTL

Protein Complex Analysis Project (PCAP): Survey of Large Protein Complexes in *Desulfovibrio vulgaris* Reveals Unexpected Structural Diversity

Bong-Gyoon Han^{1*} (bghan@lbl.gov), Ming Dong,¹ **Mark D. Biggin**,¹ and Robert M. Glaeser¹

¹Lawrence Berkeley National Laboratory, Berkeley, Calif.

Project Goals: The 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. The quaternary structures of these multi-protein complexes will be used as templates to study proteomic network within whole cells in a native-like state by cryo-EM tomography.

Single-particle electron microscopy of multiprotein complexes purified by a tagless strategy has been used to carry out an unbiased survey of the stable, most abundant multi-protein complexes in *Desulfovibrio vulgaris* Hildenborough (DvH) that are larger than Mr ~400 k. The quaternary structures for 8 of the 16 complexes purified during this work were determined by single-particle reconstruction of negatively stained specimens, a success rate about 10 times greater than that of previous “proteomic” screens. In addition, the subunit compositions and stoichiometries of the remaining complexes were determined by biochemical methods. Our results show that the structures of large protein complexes vary to a surprising extent from one microorganism to another. Except for GroEL and the 70S ribosome, none of the 13 remaining complexes with known orthologs have quaternary structures that are fully conserved. This result indicates that the interaction interfaces within large, macromolecular complexes are much more variable than has generally been appreciated. As a consequence, we suggest that relying solely on quaternary structures for homologous proteins may not be sufficient to properly understand their role in another cell of interest.

The diversity of subunit stoichiometries and quaternary structures of multiprotein complexes that has been observed in our experiments with DvH is relevant to understand-

ing how different bacteria optimize the kinetics and performance of their respective biochemical networks. It is further anticipated that imaging the spatial locations of such complexes, through the analysis of tomographic reconstructions (Downing et al.), may also be important for accurate computational modeling of such networks. While templates for some multi-protein complexes such as the ribosome or GroEL could be derived from previously determined structures, it is quite clear that single-particle electron microscopy should be used to establish the sizes and shapes of the actual complexes that exist in a new organism of interest. To not do so would be to risk searching for instances of a specific complex and finding none of them, simply because one had been searching with an invalid template.

GTL

Protein Complex Analysis Project (PCAP): Introduction of Iterative MS/MS Acquisition (IMMA) to the MALDI LC MS/MS Workflow To Enable High Throughput Protein Complex Identification using Tagless Strategy

Haichuan Liu,^{1*} Lee Yang,² Nikita Khainovski,² Ming Dong,² Evelin D. Szakal,¹ Megan Choi,² Simon Allen,¹ Terry C. Hazen,¹ Jil T. Geller,¹ Mary E. Singer,¹ Peter Walian,¹ Bing Jap,¹ Steven C. Hall,¹ Susan J. Fisher,^{1,2} H. Ewa Witkowska,¹ Jian Jin,² and **Mark D. Biggin**² (mdbiggin@lbl.gov)

¹University of California, San Francisco, Calif.; and

²Lawrence Berkeley National Laboratory, Berkeley, Calif.

Project Goals: Development and application of an intelligent method for MS/MS precursor ion selection is described in the context of the “tagless” strategy of protein complex identification. The Iterative MS/MS Acquisition (IMMA) algorithm allows for a step-wise execution of the MS/MS acquisition routine, where each step employs an automatically generated and updated exclusion list aimed at eliminating repetitive analyses of precursors derived from the previously identified polypeptides. A two-fold reduction in MS/MS acquisition time was achieved using this IMMA approach while increasing a number of polypeptides identified on the basis of 3 or more peptides by 25%.

The Protein Complex Analysis Project (PCAP) pursues two goals: (i) the identification of protein complexes in *Desulfovibrio vulgaris* Hildenborough (DvH) in order to model stress responses relevant to the detoxification of metal and radionuclide contaminated sites and (ii) the development of workflows to enable high throughput analysis of protein-protein interactions in many other prokaryotes of interest to DOE. Protein complex identification is being performed by using two complementary approaches: the well established tandem affinity purification (TAP) and a novel “tagless”

strategy. The latter approach requires MS/MS-based protein identification from a large number of protein fractions (~30,000) that result from a four-step orthogonal separation of the entire *DvH* proteome under native conditions. Putative protein complexes are identified as a collection of polypeptides that co-elute in the final protein separation step, typically size exclusion chromatography (SEC). Polypeptide elution is monitored by iTRAQ reagent-based quantitation of changes in relative concentrations of each polypeptide as it migrates through the column.

Under the current experimental design, a single pass through the whole interactome employs ~5000 mass spectrometric samples, each containing an iTRAQ octaplex. The large number of MS/MS acquisitions required to fully characterize protein-protein interactions in a single organism calls for significant enhancements in throughput at various stages of analysis, of which MS/MS analysis is one of the main rate limiting steps.

There is a significant overlap in polypeptide content among neighboring fractions due to the limited resolution of protein chromatographic steps. For this reason, we have initially opted for LC MALDI MS/MS to execute the tagless strategy workflow because of the archival capability of this approach. Our plan has been to develop a precursor ion selection algorithm that would take advantage of the information that has already been gathered from previous MS/MS analysis. Here we introduce the concept of Iterative MS/MS Acquisition (IMMA) that forms the core of an overall software development effort geared towards substantially increasing throughput of the MS/MS data acquisition.

IMMA is an information-dependent precursor selection algorithm capable of eliminating the majority of redundant acquisitions triggered by precursors representing proteins already identified. This task is accomplished by generating exclusion lists of “undesired” precursors which, due to their relatively high intensity, would otherwise be selected for analysis by the data-dependent precursor selection software provided by the manufacturer of mass spectrometer. Two parameters characteristic of a peptide, molecular mass (m/z) and retention time (RT), are utilized to generate exclusion lists. Both parameters are modeled on the basis of the experimental data. Exclusion list candidates are limited to those peptides belonging to a proteotypic category, *i.e.*, those that are likely to be preferentially detected in the course of MALDI analysis (1). In addition, a peptide retention time (RT) prediction algorithm employs RT indices for amino acids and a set of modifications that we have frequently encountered in a training set. The RT prediction algorithm also allows for alignment of each newly performed LC run to a “virtual plate”, *i.e.*, a predicted peptide retention time distribution that is built using training set data where a singular value decomposition algorithm is used to derive offsets for RT parameters. In addition to predicted proteotypic peptides, molecular ions already analyzed, satellite ions of high intensity precursors and contaminant ions are also excluded. In a typical experiment, one precursor is selected per spot at a time. The exclusion list is updated following each round of acquisition to reflect identification of new

polypeptides and hence to eliminate an additional set of redundant precursors from further consideration. At present, execution of the IMMA algorithm allows for a two-fold reduction of acquisition time while using acceptance criteria of at least 3 confidently identified peptides per polypeptide. At the same time, the number of confidently identified polypeptides increased by 25%. Efforts are underway to extend the IMMA concept to multiple sets of LC MALDI runs, starting with all fractions from a single SEC column. In this scenario, the reference database of the already identified peptides will incorporate all species observed to date. Exclusion lists will be refined by using the exact rather than predicted retention times for species that were observed in earlier analyses. In addition, inclusion lists based upon peptides confidently identified in previous analyses will be executed to enable targeted monitoring of elution profiles of polypeptides identified in the same protein separation space.

We acknowledge members of Ruedi Aebersold's group for providing us with the “PeptideSieve” software for generating proteotypic peptides for *DvH* proteome according to their published algorithm: Mallick et al., (2007) Nat. Biotech. 25:125-131.

GTL

Protein Complex Analysis Project (PCAP): Isolation and Identification of Membrane Protein Complexes from *D. vulgaris*

Peter J. Walian^{1*} (PJWalian@lbl.gov), Simon Allen,² Lucy Zeng,¹ Evelin Szakal,² Eric Johansen,² Steven C. Hall,² Susan J. Fisher,^{1,2} Mary E. Singer,¹ Chansu Park,¹ Terry C. Hazen,¹ H. Ewa Witkowska,² **Mark D. Biggin**,¹ and Bing K. Jap¹

¹Lawrence Berkeley National Laboratory, Berkeley, Calif.;
²University of California, San Francisco, Calif.

Project Goals: As a component of the Protein Complex Analysis Project (PCAP) our goals are to 1) develop a pipeline for high-throughput isolation and identification of *Desulfovibrio vulgaris* membrane protein complexes and 2) use this methodology for identifying complexes isolated from wild type cells grown under normal conditions and to characterize changes in these complexes in response to environmentally relevant stressors.

This component of the Protein Complex Analysis Project (PCAP) has been focused on optimization of a processing pipeline to isolate membrane protein complexes from *D. vulgaris*, and identify their subunits by mass spectrometry. As part of PCAP's effort to model stress responses relevant to the detoxification of metal and radionuclide contaminated sites, we are applying a suite of the methods thus developed to catalog, as completely as possible, stable membrane protein complexes present in wild type cells grown under standard conditions as well as in the presence of stressors (*e.g.*, nitrate and sodium chloride). Our ultimate goal is to characterize changes in the relative abundance,

composition and stoichiometry of membrane protein complexes in cells with perturbed stress response pathways.

Membrane protein complexes are particularly challenging to purify and characterize. Largely due to the requirement for detergent solubilization, stable isolation of homogeneous intact membrane protein complexes in detergents typically requires separation conditions that are different from those used for water soluble proteins. Therefore, we are developing a novel “tagless” strategy specifically optimized for purifying membrane proteins and then identifying them by mass spectrometry.

In this approach, isolated *D. vulgaris* cell membranes are sequentially processed, initially using mild detergents suited for the solubilization of inner membrane proteins (such as Triton X-100 and C12E9) and then with a second more effective detergent (*e.g.*, octyl POE or octyl glucoside) to solubilize proteins of the outer membrane. Isolation of tagless complexes expressed in lower copy numbers is especially demanding and has driven us to process increasingly larger amounts of membrane (derived from cells obtained from over 100 liters of cell culture) in a given purification. To purify candidate complexes of the inner- and outer-membranes, ion exchange (IEX) and molecular sieve chromatography have been used. Fractions obtained from these procedures are further analyzed using SDS and blue native gel electrophoresis to isolate candidate complexes, obtain molecular weight estimates and to prepare complex subunit samples for mass spectrometry (MS) analysis.

To optimize chromatographic resolution and sample management under these conditions, proteins bound to ion exchange media are eluted using a fine step gradient profile. Following ion exchange chromatography, the collected fractions of each elution peak are concentrated and applied to a molecular sieve column. Molecular sieve column elution fractions are subjected to a combination of blue native and SDS PAGE. Proteins of potential complexes are extracted from the bands of blue native PAGE gels and directly transferred to a second dimension SDS PAGE. To improve upon extraction efficiency and the rate of sample production for mass spectrometry (MS), an alternative approach of gel-to-gel protein transfer has been employed in the past year. In this procedure, length-wise sections of blue-native PAGE gel lanes are being placed directly upon the stacking sections of denaturing gels and subjected to a second dimension of SDS PAGE. Protein bands or spots excised from these gels are subjected to in-gel digestion and analysis by liquid chromatography electrospray ionization tandem mass spectrometry (LC-ESI-MS/MS) and consequently identification by searching a custom *D. vulgaris* database using the Mascot search engine.

Using both native gel single-band extractions and two-dimensional techniques, over fifty membrane protein complexes have been identified. About one-fourth of these complexes are heteromeric.

We have recently begun preliminary analysis of *D. vulgaris* protein complexes from stressed cell cultures. Stressed cell conditions surveyed were stationary growth, elevated salt

and elevated nitrate. While changes to proteins of the inner membrane, as assessed by SDS PAGE of IEX fractions, appeared to be relatively modest, changes to proteins of the outer membrane appear to be more pronounced. Changes to proteins of the outer membrane, in response to stressors, can be expected given that these proteins represent the first line of defense against environmental changes.

—
GTL

Protein Complex Analysis Project (PCAP): Reconstruction of Multiple Structural Conformations of Macromolecular Complexes Studied by Single-Particle Electron Microscopy

Maxim Shatsky^{1,2*} (maxshats@compbio.berkeley.edu),
Richard J. Hall,² Jitendra Malik,² and Steven E. Brenner^{1,2}

¹Lawrence Berkeley National Laboratory, Berkeley, Calif.;
and ²University of California, Berkeley, Calif.

In the hierarchy of biological visualization, 3D electron microscopy (EM) bridges the gap between the object sizes studied by X-ray crystallography and light microscopy. Single-particle EM is routinely used to resolve the three-dimensional structure of large macromolecular assemblies. As a part of LBNL's GTL PCAP project, single-particle EM molecular imaging is being applied to study protein complexes of *Desulfovibrio vulgaris*. These studies analyze stoichiometry and structural changes of the protein complexes under physiological and stress conditions. The structures obtained from EM experiments will serve as templates for identification and localization of the macromolecules in the images of the complete bacterial cells produced by the electron tomography. This structural information will facilitate the understanding of the functional roles of protein complexes in bioremediation studies by *D. vulgaris*.

One of the major challenges in single-particle EM is structural heterogeneity of the studied particles. Particles can adopt different conformations and can be found in assemblies with alternative quaternary-structure. Computational methods for image processing and three-dimensional structure determination play a crucial role in single-particle EM. Most commonly used computational approaches assume that the imaged particles have homogeneous shape and quaternary-structure. When this assumption is violated, the product of the three-dimensional reconstruction is one low resolution structure. Our aim is to improve and redesign various computational stages of particle reconstruction, taking into account the heterogeneity of the data.

We present a new computational method that reveals the existence of different conformational states of the studied macromolecular complexes. It is able to automatically classify the experimental images into homogeneous subsets which produce structurally different models. The method achieves high accuracy on synthetic data sets and shows

promising results on real data. In one test case, we successfully differentiated between the real experimental images of human translation initiation factor eIF3 and of eIF3 complexed with hepatitis C virus (HCV) IRES RNA. In another experiment, where we used experimental images of human RNA polymerase II, our method produced two models that show a substantial conformational flexibility of the protein complex. Our next goal is to apply this approach on various macromolecular complexes studied in PCAP. We expect that under some stress conditions protein complexes may be found in multiple structural configurations. Recognition of various structural conformations and, via electron tomography, their cell localization will aid in understanding of the *D. vulgaris* response to stress conditions related to bioremediation.

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 Jason Baumohl,¹ Keith Keller,¹ and **John-Marc Chandonia**^{1,2*} (jmchandonia@lbl.gov)

¹Lawrence Berkeley National Laboratory, Berkeley, Calif.; and ²University of California, Berkeley, Calif.

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.

We have developed 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. In the last year, we released 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, and was used to build much of the current PCAP LIMS. WIST is available for download at <http://vimss.sourceforge.net>, under an open source license.

We have compared complexes identified in the PCAP tagless purification pipeline to complexes formed by orthologous proteins in other bacteria. We have discovered a surprisingly low degree of conservation in the stoichiometry of such complexes. This finding is consistent with the low degree of structural conservation that we observed in a smaller sampling of complexes studied using single particle EM (see the PCAP EM poster for details). We are currently analyzing results for outer membrane complexes under several stress conditions, and comparing data on the composition of complexes purified by the tagless approach to data on complexes isolated using affinity tags. We are also working with the Environmental Stress Pathway Project (ESPP) to integrate data on protein complexes into the MicrobesOnline website (<http://microbesonline.org>) for public dissemination.

GTL

Protein Complex Analysis Project (PCAP): Towards Localization of Functionality in *Desulfovibrio vulgaris* by Electron Microscopy

David A. Ball* (DABall@lbl.gov), Jonathan Remis, Manfred Auer, and **Kenneth H. Downing**

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

Project Goals: We aim to demonstrate the feasibility of using electron microscopy for high-throughput structural characterization of multi-subunit complexes in microbes. We will determine the spatial organization and relative locations of complexes within individual cells and map the relation between particular enzymes and other molecules and sites of metal reduction.

The anaerobic sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough (DvH) is used as a model organism for the study of environmental bioremediation of heavy metal and radionuclide contamination. DvH has the ability to efficiently reduce toxic heavy metals such as uranium and chromium and is of particular interest to DOE for use in high-risk metal contaminated sites. Electron microscopy is being used in several ways to gain insights on the molecular pathways involved in these and other processes.

Cryo-electron tomography and high-resolution single particle analysis (see abstract by BG Han et al., these proceedings) have been used to study the morphology of intact DvH cells and the larger protein complexes contained within them. During this specific study a number of surprising phenotypes and anomalies have been observed. The cytoplasm of DvH cells appears significantly more opaque than that of other bacterial species of similar size and doubling time, making it more difficult to identify molecular features. In cells that have been grown in sulfate up to and past cessation of growth, large internal "balls" have been observed, identified as sulfur deposits using Energy Dispersive X-ray Analysis. These large 'sulfurous balls' increase in

both size and number over time culminating in their occurrence across the total DvH population. Interestingly alongside the appearance of these electron dense sulfurous balls we find the emergence of a secondary population of rod-shaped lipid structures or “poles”. These poles also appear to increase in number and regularity over time, are very beam sensitive and are present in two distinct length classes.

In addition to observing cells in their planktonic state, we have also examined DvH that were allowed to form a community (biofilm), which is more similar to their natural environmental habitat. Using high-pressure freezing and freeze-substitution or microwave-assisted processing, followed by resin-embedding, serial sectioning and electron tomography, we have identified a number of biofilm-specific structures including extracellular filaments and vesicle-like structures that assemble into clusters not unlike grapes on a vine, which are associated with extracellular metal deposition. Metal deposition was found to be non-uniform, with some cells appearing particularly active and others devoid of metal deposits. In addition to metal deposits near the bacterial cell surface, we have observed long-range sheet-like metal deposits that can extend for tens if not hundreds of microns. Based on their staining pattern we believe that these sheets contain a lipid core that is surrounded by metal deposits.

Current work is aimed at further characterizing localization of intra- and extracellular protein complexes that are involved in the cell's various metabolic pathways.

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}

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

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 knockout mutation strains are being produced to determine the effect of eliminating expression of components of putative stress response protein complexes. Both types of engineered strains are being generated using a two-step procedure that first integrates and then cures much of the recombinant DNA from the endogenous chromosomal location of the target gene. We are developing new counter selective markers for *D. vulgaris*. This procedure will; 1) allow multiple mutations to be introduced sequentially; 2) facilitate the construction of in-frame deletions; and 3) prevent polarity in operons. The Microbiology Subproject provides high throughput phenotyping of all engineered strains to determine if any show phenotypic changes. We also determine if the tagged proteins remain functional and that they do not significantly affect cell growth or behavior. The knockout mutations are tested in a comprehensive set of conditions to determine their ability to respond to stress. High throughput optimization of culturing and harvesting of wild type cells and all engineered strains are used to determine the optimal time points, best culture techniques, and best techniques for harvesting cultures using real-time analyses with synchrotron FTIR spectromicroscopy, and other methods. Finally, we are producing large quantities of cells under different conditions and harvesting techniques for 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 1,100 individual productions. Production volumes range from less than 10 mL 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.

During the past year we have continued to evaluate three methods for genetic construction of tagged genes for TAP and visualization. These included the commercially available Gateway system and recombineering protocol against the recently developed SLiC (Sequence and Ligation independent Cloning) method. Further, we have continued to streamline the high-throughput pipeline for clone construction and implement QA/QC and verification protocols that are amenable to high-throughput. Early results suggest that the SLiC method is not only cheaper, faster and can tag any gene regardless of placement within the operon, but also yields a greater pipeline efficiency of up to 30% (verified construct/electroporation). Our current throughput level with ~30 electroporations per week yields ~5 confirmed constructs (16%) with the Gateway system with a turnover time of ~6 weeks from gene amplification through to a verified tag construct.

Using a commercial liquid handling system (Biomek NX) and in-house developed LIMS system for sequence analyses and sample tracking, we processed 576 genes from the *D. vulgaris* genome for tagged construct generation using the commercial TOPO-Gateway (R) strategy. Of these we were able to successfully confirm 360 constructs bearing the STF/SPA and SNAP tags. The tagged constructs were then fed into the electroporation pipeline for tagged strain construction based on the single cross-over chromosomal integration approach which is limited to genes located on terminal ends of their operons. For tagging genes located within operons using the double-recombination approach, we developed a novel scheme based on the Sequence and Ligation Independent (SLiC) cloning technique as detailed in the adjacent poster.

The Center for Molecular and Cellular Systems

GTL

An Integrative Strategy for the Determination of the Modular Structure of Functional Networks of *Rhodospseudomonas palustris*

William R. Cannon^{2*} (William.Cannon@pnl.gov), Don Daly,² Mudita Singhal,² Lee Ann McCue,² Ronald Taylor,² Dale A. Pelletier,¹ Gregory B. Hurst,¹ Denise D. Schmoyer,¹ Jennifer L. Morrell-Falvey,¹ Brian S. Hooker,² Chongle Pan,¹ W. Hayes McDonald,^{1,3} **Michelle V. Buchanan**,¹ and H. Steven Wiley²

¹Oak Ridge National Laboratory, Oak Ridge, Tenn.; ²Pacific Northwest National Laboratory, Richland, Wash.; and ³Vanderbilt University, Nashville, Tenn.

Project Goals: The overall goal of the Center for Molecular and Cellular Systems (CMCS) is to provide a capability for generating high quality protein-protein interaction data from a variety of energy- and environment-relevant microbial species.

Protein-protein interactions were used as the foundation for an integrative approach for determining the modular structure of *Rhodospseudomonas palustris* cellular networks. *R. palustris*, a metabolically versatile anoxygenic phototrophic bacterium, is the current target for the Genomics:GTL Center for Molecular and Cellular Systems (CMCS). Our 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 developed an approach where multiple sources of information are integrated to provide a more comprehensive perspective of the cellular networks than can be provided by protein interactions alone. The interaction data are integrated with functional clues from operon structure and with manual and automated analysis of functional subsystems from molecular machines and cellular processes. Transcriptional regulatory elements are then overlaid on the protein-interaction maps to provide an integrative perspective on the cellular machines.

Using this approach, we have reconstructed a wide ranging, catalogue of protein complexes and interactions involved in a diverse and rich set of cellular processes including nitrogen fixation, electron transfer, photosynthesis, protein synthesis, ATP synthesis, central metabolism, fatty acid synthesis and uncharacterized processes.

Analysis of the functional network associated with nitrogen fixation identified a number of hubs and bottlenecks. Of the five proteins that are both hubs and bottlenecks, the three proteins that have functional annotations are all involved in electron transfer or synthesis of electron transfer proteins.

The two unknown proteins, in addition, appear to also be related to electron transfer processes via their interaction partners. We have also compared the global interaction network and individual subnetworks with those inferred from protein interaction data gathered in other bacteria, such as *E. coli* and the nitrogen-fixing soil bacterium *Mesorhizobium loti*. Of the comparable interactions studied between *E. coli* and *R. palustris*, 20% are shared between the two species. Our comparative approach both improves the functional annotation of interaction networks of non-model species, and reveals fundamental architectural principles of the biochemical networks of microbes.

GTL

An Imaging-Based Assay with High Sensitivity for Confirming and Characterizing Protein Interactions

Jennifer L. Morrell-Falvey* (morrelljl1@ornl.gov), A. Nicole Edwards, Jason D. Fowlkes, Robert F. Standaert, Dale A. Pelletier, Mitchel J. Doktycz, and **Michelle V. Buchanan**

Oak Ridge National Laboratory, Oak Ridge, Tenn.

Project Goals: The overall goal of the Center for Molecular and Cellular Systems (CMCS) is to provide a capability for generating high quality protein-protein interaction data from a variety of energy- and environment-relevant microbial species.

Identifying and characterizing protein interactions are essential for understanding and modeling cellular networks. Several methods exist to assay protein interactions; however, none are known to provide both confirmation of protein interactions and simultaneous quantification of biophysical parameters (binding strengths and association/dissociation rates) *in vivo*. We are currently developing an approach that combines an imaging-based protein interaction assay with a fluorescence photobleaching and recovery technique (iFRAP), and computer simulations to provide a facile, general method for quantifying protein binding affinities *in vivo*. This protein interaction assay relies on the co-localization of two proteins of interest fused to DivIVA, a cell division protein from *Bacillus subtilis*, and green fluorescent protein (GFP). We have modified this imaging-based assay to facilitate high-throughput applications by constructing new vectors encoding N- and C-terminal DivIVA or GFP molecular tag fusions based on site-specific recombination technology and have determined the range of binding affinities that can be detected using this assay. The sensitivity of the assay was defined using a well-characterized protein interaction system involving the eukaryotic nuclear import receptor subunit, Importin α (Imp α) and variant nuclear localization signals (NLS) representing a range of binding affinities. Using this system, we demonstrate that the modified co-localization assay is sensitive enough to detect protein interactions with K_d values that span over four

orders of magnitude (1nM to 15 μ M). Moreover, the spatial confinement of the interacting proteins should also enable measurements of binding constants *in vivo* using iFRAP. Initial experiments demonstrate the anticipated decay of fluorescence at the cell poles indicative of a binding interaction. The statistical variance is reported as a function of K_d .

GTL

Functional Characterization of Protein Complexes and Cellular Systems in *Rhodopseudomonas palustris* using Stable Isotopic Labeling and Quantitative Proteomics

Chongle Pan^{1*} (panc@ornl.gov), W. Judson Hervey IV,^{1,2} Michael S. Allen,^{1,3} Dale A. Pelletier,¹ Gregory B. Hurst,¹ and **Michelle V. Buchanan**¹

¹Oak Ridge National Laboratory, Oak Ridge, Tenn.;

²Graduate School of Genome Science and Technology, University of Tennessee-Oak Ridge National Laboratory, Oak Ridge, Tenn.; and ³University of North Texas, Denton, Tex.

Project Goals: The overall goal of the Center for Molecular and Cellular Systems (CMCS) is to provide a capability for generating high quality protein-protein interaction data from a variety of energy- and environment-relevant microbial species.

Responses to various types of stresses can be identified in microbial cells at the protein level using quantitative comparative proteomics. Proteins can be metabolically labeled with stable isotopes, such as ¹⁵N and ¹³C, by growing a microorganism of interest in a medium enriched in those stable isotopes. Stable isotope labeling enables large-scale accurate protein quantification by mass spectrometry, referred to as quantitative proteomics. Here, quantitative proteomics was used to characterize protein complexes and cellular systems in *Rhodopseudomonas palustris*.

Stable isotope labeling can be used to evaluate performance of affinity-tagging strategies for studies of protein-protein interactions both at the level of the protein complex, and at the level of the proteome. Affinity-purified protein complexes are often accompanied by background, non-specific proteins. In this study, authentic interacting proteins of a model complex, DNA-dependent RNA polymerase (RNAP), were successfully distinguished from artificial co-isolating proteins by the isotopic differentiation of interactions as random or targeted (I-DIRT) method (A. J. Tackett et al. J. Proteome Res. 2005, 4 (5), 1752-1756). To investigate broader effects of bait protein production on bacterial metabolism, we compared proteomes from strains harboring the plasmid that encodes an affinity-tagged subunit (RpoA) of the RNAP complex with the corresponding wild-type strains using stable isotope metabolic labeling. Expression of plasmid-encoded bait protein significantly

induced the expression of several proteins involved in amino acid biosynthesis.

Cellular systems function not only via a physical interaction network but also within a regulatory network. In results from the Center for Molecular and Cellular Systems, observations of protein-protein interactions among a putative anti- σ factor RPA4224, an extracytoplasmic function (ECF) σ factor RPA4225, and the predicted response regulator RPA4223 led us to study this system further. We characterized a global stress regulon controlled by RPA4225 in *R. palustris* using quantitative proteomics. Changes in expression of several genes resulting from overproduction of RPA4225 were further verified by quantitative PCR. Furthermore, most of the strongly up-regulated proteins revealed a conserved binding motif, which we also found in the promoters of over 150 genes, including general stress proteins. These data suggest that RPA4225 controls a global stress regulon that may be conserved among several members of α -Proteobacteria.

These studies showcased the biological insights one can obtain using stable isotope labeling and quantitative proteomics. In addition, new methods based on stable isotope labeling are under development at the Center for Molecular and Cellular Systems for protein absolute quantification and complex dissociation kinetics.

—
GTL

Protein-Protein Interactions in *Rhodopseudomonas palustris* at the Genomics:GTL Center for Molecular and Cellular Systems

Dale A. Pelletier^{1*} (pelletierda@ornl.gov), Kevin K. Anderson,² William R. Cannon,² Don S. Daly,² Brian S. Hooker,² H. Steven Wiley,² Lee Ann McCue,² Chongle Pan,¹ Manesh B. Shah,¹ W. Hayes McDonald,¹ Keiji G. Asano,¹ Gregory B. Hurst,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 and **Michelle V. Buchanan**¹

¹Oak Ridge National Laboratory, Oak Ridge, Tenn.; and ²Pacific Northwest National Laboratory, Richland, Wash.

Project Goals: The overall goal of the Center for Molecular and Cellular Systems (CMCS) is to provide a capability for generating high quality protein-protein interaction data from a variety of energy- and environment-relevant microbial species.

The overall goal of the Center for Molecular and Cellular Systems (CMCS) is to provide a capability for generating high quality protein-protein interaction data from a variety of energy- and environment-relevant microbial species. The CMCS approach combines expression of affinity tagged proteins, affinity purification of interacting proteins, and

tandem mass spectrometric identification of these proteins (Pelletier et al., *J. Proteome Research* 2008, 7, 3319-3328).

We have recently completed the characterization of soluble protein-protein interactions in *Rhodopseudomonas palustris*. These results are the first large-scale protein-protein interaction results of this type in an organism other than a model system such as *E. coli* or yeast. The protein-protein interactions from the metabolically versatile *R. palustris* provide insights into microbial processes of high relevance to DOE missions, including the ability to produce hydrogen, to degrade lignin monomers, to perform photosynthesis, and to fix nitrogen.

As of early December 2008, 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. Over 800 affinity-tagged bait proteins have been expressed, affinity purified, and subjected to mass spectrometry (MS) analysis to identify interacting proteins. Criteria for choosing these bait proteins from among the >4800 in the *R. palustris* predicted proteome included predicted location in the cytosol, and previous detection at medium to high levels in proteomics measurements (VerBerkmoes et al., *J. Proteome Research* 2006, 5, 287-298). Quantitative estimates of confidence in putative bait-prey interactions identified from the MS analysis are obtained using the statistical tool BePro3 (Sharp et al., *J. Proteome Research* 2007, 6, 3788-3795). From thousands of putative interactions, a few hundred survive as candidates for further study, based on preliminarily chosen BePro3 threshold values. Integration with additional data, including comparative genomics, transcriptomics, operon structure, and regulatory networks, that aid in interpretation and functional annotation of novel interactions is described in the poster "An Integrative Strategy for the Determination of the Modular Structure of Functional Networks of *Rhodopseudomonas palustris*" by Cannon *et al.* Results of the protein-protein interaction survey in *R. palustris* are available through the publicly accessible Microbial Protein-Protein Interaction Database (MiPPI.ornl.gov).

Validation of Genome Sequence Annotation

GTL

GTL

Toward a High Throughput Functional Annotation Pipeline for Fungal Glycoside Hydrolases

Scott E. Baker^{1*} (scott.baker@pnl.gov), Ellen A. Panisko,¹ Deanna Auberry,¹ Beth Hofstad,¹ Jon K. Magnuson,¹ Barbara Robbertse,² Adrian Tsang,³ and Frank Collart⁴

¹Chemical and Biological Process Development Group, Pacific Northwest National Laboratory, Richland, Wash.; ²Oregon State University, Corvallis, Ore.; ³Concordia University, Montreal, Canada; and ⁴Argonne National Laboratory, Argonne, Ill.

Project Goals: Our goal is to characterize glycoside hydrolases using an approach that is based on bioinformatic and proteomic analyses, high throughput protein expression and enzymatic activity assays.

Glycoside hydrolases (GHs) from fungi are a key component of the biological disassembly of cellulosic biomass into its component sugar monomers. In the DOE vision of the biorefinery, these sugars are subsequently converted to fuels and chemical products by microbial bioprocesses. As the number of fungal genome sequences increases, so too does the number of GHs that lack any functional characterization. We have initiated an approach to GH characterization that is based on bioinformatic and proteomic analyses, high throughput protein expression and enzymatic activity assays. In order to assess the diversity of an important class of GHs we performed a phylogenetic analysis of GH family 7. Our analysis indicates that GH7s have been lost or duplicated numerous times across the fungal kingdom. Additionally, our results show that encoded within the genome of a many individual fungal species, there is a tremendous range of amino acid sequence diversity that may be indicative of the variety of environmental growth conditions encountered by these organisms.

We are characterizing GHs within a variety of individual fungal species. Initial studies are bioinformatic and proteomic in nature. These are followed by high throughput heterologous protein production and enzymatic activity assays of individual GHs. *Aspergillus niger*, a filamentous ascomycete fungus, was the initial organism chosen for GH characterization in our pipeline. Current organism targets undergoing proteomic analyses include the basidiomycete brownrot, *Postia placenta*, a close relative of *A. niger*, *Aspergillus carbonarius*, the zygomycete, *Phycomyces blakesleeana* and a thermophilic ascomycete fungus, *Thielavia terrestris*.

Annotation of the *Clostridium phytofermentans* ORFome by Proteogenomic Mapping

Andrew Tolonen*

Dept. of Genetics, Harvard Medical School, Boston, Mass.

Project Goals: The goal of this project is to analyze the *Clostridium phytofermentans* ORFome by building a proteogenomic map from data obtained by mass spectrometry.

Clostridium phytofermentans is gram positive anaerobe that efficiently converts the two most abundant constituents of plant feedstocks, cellulose and hemicellulose, to ethanol and hydrogen. The breakdown of cellulosic biomass is accomplished by a diverse set of saccharolytic enzymes; the resulting hexoses and pentoses are then fermented, with ethanol as the primary product. *C. phytofermentans* is thus a model system for the direct conversion of cellulosic feedstocks to biofuels. The *C. phytofermentans* genome was sequenced by the DOE JGI. The genome is 4.8 Mb and contains 3926 candidate protein-encoding gene models. In this project we validated and corrected the JGI gene annotations by building a proteogenomic map from data obtained by mass spectrometry.

A proteogenomic map is an application of proteomics to genome annotation: novel proteins not included in the standard annotation are identified, the boundaries of annotated ORFs are defined, and ORFs that are considered 'hypothetical' based upon computational gene prediction are verified. Proteogenomic mapping is an ideal complement to computational genome annotation because mass spectrometry provides direct, molecular evidence about which ORFs are translated into proteins. In this study, we analyzed the proteome of *C. phytofermentans* growing on different carbon sources: glucose, cellulose, and hemicellulose using an LTQ FT Ultra Hybrid mass spectrometer. Greater than 10,000 unique peptides were mapped to the *C. phytofermentans* genome to create a proteogenomic map. This presentation will compare features of the computational genome annotation of *C. phytofermentans* to the genome annotation from the proteogenomic map.

Protein Functional Assignment Using a Fluorescence-Based Thermal Shift Assay

Ashley M. Frank^{1*} (afrank@anl.gov), Sarah E. Giuliani,¹ William Studier,² and Frank Collart¹

¹Biosciences Division, Argonne National Laboratory, Lemont, Ill.; and ²Biology Division, Brookhaven National Laboratory, Upton, N.Y.

Project Goals: Generate specific functional assignments for selected protein families such as ABC transporters, transcription factors and two component systems. Assemble an integrated data set that enables identification of sequence motifs associated with biological function of specific protein families.

We are developing *in vitro* methods for functional characterization of proteins that can be used to interrogate proteins involved in cellular metabolic, sensory and response pathways. Our approach for the identification of bound ligands and assignment of function uses a fluorescence-based thermal shift (FTS) assay for high-throughput screening of protein-ligand interactions. The FTS approach is a target-independent assay that uses a fluorescent dye to monitor protein unfolding. 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- or 384-well plate format.

Our initial study focused on solute-binding proteins in the bacterial ABC transporter family. These transporters are essential membrane transport components in many organisms and transport a diverse range of ligands, but a specific functional role has not been assigned for a majority of these proteins in most organisms. The assay was validated with a set of six proteins with known binding specificity and was consistently able to map proteins with their known binding ligands. The assay also identified additional candidate binding ligands for several of the amino acid binding proteins in the validation set.

We extended this approach to additional targets and demonstrated the ability of the FTS assay to unambiguously identify preferential binding for several homologs of amino acid binding proteins with known specificity and to functionally annotate a protein of unknown binding specificity. The targets in this evaluation set had various degrees of experimental functional characterization (Table 1) but none of the individual proteins were represented as structures in the PDB. In all cases, the assay predicted ligand assignments that were consistent with ligand assignments inferred by other experimental approaches.

The FTS approach was also applied to a set of targets from *S. oneidensis* which are clustered into two COGs (Table 1). Five of the *Shewanella* proteins are categorized into COG083ET, a cluster which contains amino acid binding and signal transduction proteins. These five proteins

were annotated as periplasmic binding proteins but only one (NP_716672) was specifically annotated as an amino acid-binding protein. The remaining protein is grouped in COG2998H, a cluster with the description of “ABC-type tungstate transport system, permease component,” but is annotated as a hypothetical protein (NCBI database). The *S. oneidensis* target selected from COG2998H was strongly stabilized by tungstate and showed some stabilization with molybdate. This protein has a signal peptide and is part of an operon encoding an ABC-type transporter and ATPase. Specific binding to arginine was detected for a different protein (NP_716672) which was originally annotated as an amino acid binding protein (Table 1).

Table 1. FTS Assay Results for Solute-Binding ABC-type Transporter Proteins Varying in Degree of Experimental Functional Annotation

Protein	Source Organism	Accession #	Tm (°C) (No Ligand)	Binding Ligand(s)	Tm Shift (°C) 1000µM(°C)	Tm Shift (°C) 200µM(°C)	Tm Shift (°C) 20µM(°C)
Highly characterized	<i>Escherichia coli</i>	NP_410281	55.0	ARG	5.0	5.0	0.0
DL-methionine transporters (yubA)	<i>Escherichia coli</i>	NP_417025	55.5	MET	5.0	4.0	1.0
Lactose, arginine, or citrate binding protein	<i>Escherichia coli</i>	NP_410511	45.5	ARG, ORN, CIT	3.0, 7.0, 5.0	5.0, 4.0, 3.0	1.0, 1.0, 0.0
COG084ET protein	<i>Shewanella oneidensis</i> M91	NP_716677	52.5	ARG	11.0	7.0	7.0
COG084ET protein	<i>Shewanella oneidensis</i> M91	NP_717122	55.5	Met	—	—	—
COG084ET protein	<i>Shewanella oneidensis</i> M91	NP_717449	55.5	Met	—	—	—
COG084ET protein	<i>Shewanella oneidensis</i> M91	NP_716950	57.5	Met	—	—	—
COG084ET protein	<i>Shewanella oneidensis</i> M91	NP_716950	53.5	Met	—	—	—
COG083H protein	<i>Shewanella oneidensis</i> M91	NP_702251	51.5	Tungstate, Molybdate	20.0, 4.0	—	—

The assay is implemented in a microwell plate format and provides a rapid approach to validate an anticipated function or to screen proteins of unknown function. The ABC-type transporters family is ubiquitous and transports a variety of biological compounds, but the current annotation of the ligand binding proteins is limited to mostly generic descriptions of function. The results illustrate the feasibility of the FTS assay to improve the functional annotation of binding proteins associated with ABC-type transporters and suggest this approach that can also be extended to other proteins families.

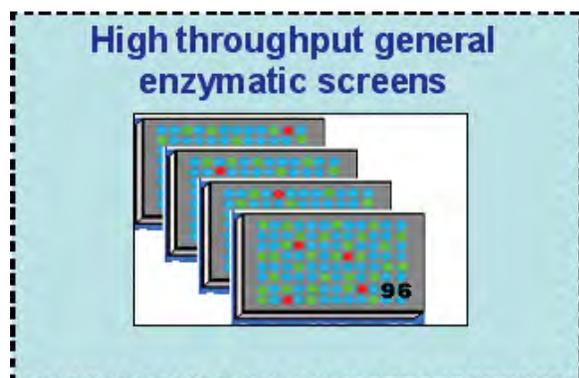
Assignment of Enzymatic Function for Core Metabolic Enzymes

Elizabeth Landorf,^{1*} Vincent Lu,¹ Gopi Podila,² Michael Proudfoot,³ Alexander Yakunin,³ and Frank Collart¹ (fcollart@anl.gov)

¹Biosciences Division, Argonne National Laboratory, Lemont, Ill.; ²Dept. of Biological Sciences, University of Alabama, Huntsville, Ala.; and ³Banting and Best Dept. of Medical Research, University of Toronto, Toronto, Canada

Project Goals: Benchmark the utility of general enzymatic screens for the improvement of functional assignments. Evaluate the utility for specific assignment of function and the overall impact on the annotation set.

With over 800 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. In the first stage of the project we have used the general enzymatic screens to functionally map a subset of the conserved hypothetical proteins from *Sewanella oneidensis* which contains ~800 members. To date, we have cloned ~500 cloned targets and 313 of these targets had levels of expression and solubility. Over 300 proteins were purified under Ni-affinity chromatography and ~200 of this set have been delivered to the University of Toronto for enzymatic screening. The screening data from this initial set suggest high throughput enzymatic screens have general utility for the identification of a functional category. Approximately 10-20% of the screened targets showed some activity using the current library of chemical compounds.



A directed screening strategy process was applied to a subset of proteins that demonstrated thioesterase, esterase or HAD hydrolase activity. For the targets that tested positive for thioesterase substrates, the directed screening was able to identify specific activity profiles for acetyl CoA substrates. Similar activity and biochemical profiles could be generated for other target families identified in the general screening process.

These studies benchmark the utility of enzymatic screening for functional annotation for proteins of unknown function. These results also suggest that bioinformatic prescreening and restriction of the generalized screens to specific ligand categories can improve the specificity and overall impact of the functional assignment. This approach will provide a foundation to assess the capabilities for specific functional assignments for a substantial number of unknown prokaryotic and eukaryotic proteins.

Quality Improvement Process for JGI-ORNL Microbial Annotation Pipeline

Miriam Land^{1*} (landml@ornl.gov), Loren Hauser,¹ Janet Chang,¹ Cynthia Jefferies,¹ Gwo-Liang Chen,¹ Frank Larimer,¹ Natalia Mikhailova,² Natalia Ivanova,² Athanasios Lykidis,² Galina Ovchinnikova,² Amrita Pati,² Nikos Kyrpides,² and **Bob Cottingham**¹

¹Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; and ²DOE Joint Genome Institute, Walnut Creek, Calif.

Project Goals: With the extension of high-throughput sequencing to include microbial genomes, there has been a rapid expansion of microbial genomic data, requiring the development of comprehensive automated tools to provide in-depth annotation to keep pace with the expanding microbial dataset. We continue to develop tools for comparative multiple genome analysis that provide automated, regularly updated, comprehensive annotation of microbial genomes using consistent methodology for gene calling and feature recognition. An automated analysis pipeline provides annotation for the microbial sequencing projects being carried out at the JGI. Multiple gene-finders are used to construct a candidate gene model set. The conceptual translations of these gene models are used to generate similarity search results and protein family relationships; from these results a metabolic framework is constructed and functional roles are assigned. Simple repeats, complex repeats, tRNA genes and other structural RNA genes are also identified. Annotation summaries are made available through the JGI Microbial Sequencing web site; in addition, results are integrated into interactive display schemes at ORNL. Comprehensive representation of microbial genomes requires deeper annotation of structural features, including operon and regulon organization, promoter and ribosome binding site recognition, repressor and activator binding site calling, transcription terminators, and other functional elements. Sensor development is continuing to enhance access to these features. Linkage and integration of the gene/protein/function catalog to phylogenomic, structural, proteomic, transcriptional, and metabolic profiles are being developed. The expanding set of microbial genomes comprises an extensive resource for comparative genomics: new tools continue to be developed for rapid exploration of gene and operon phylogeny, regulatory networking, and functional proteomics.

The US DOE Joint Genome Institute (JGI) performs high-throughput sequencing and annotation of microbial genomes through the DOE Microbial Genome Program (MGP). The world-wide rate of sequencing is resulting in a rapid expansion of microbial genomic data, which requires the development of comprehensive automated tools to provide in-depth annotation which can keep pace with the expanding microbial dataset.

JGI-ORNL annotates the microbial genomes which are sequenced by JGI. We have and continue to develop tools for genome analysis that provide automated annotation of microbial genomes using consistent methodologies. One such tool is our new gene caller, Prodigal, which has its own abstract.

Unfortunately, all automated methods have inherent errors because they are based on comparisons to existing data sources which vary in quality and applicability. We have implemented processes to monitor and improve the quality of the automated annotation while making the process fully automated. This will reduce the sources of error and will provide users with qualifiers on the annotation.

The JGI annotation process includes a quality control (QC) step. The annotation is reviewed for pseudo genes, missed genes, and potential changes in start codons. Patterns identified in the QC process are fed back into the initial annotation process to improve the quality of gene predictions. Continual improvement is also made to the QC process to make it both efficient and effective.

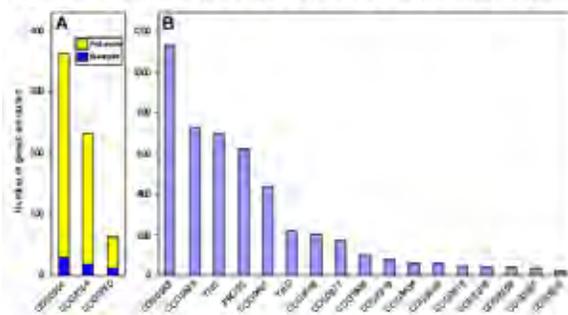
Inference flags have been added to the GenBank annotation. The inference and notes provide users with basis for determining the source and quality of the annotation.

ORNL has created a baseline database consisting of representative sequences from a taxonomically diverse group of organisms. This representative group of DNA sequences was initially annotated in early 2008. As the underlying databases change and tools are upgraded, the baseline database is reannotated and an impact analysis is done. This ensures that changes to the automated annotation have a positive effect on the outcome. In addition, genomes originally annotated by JGI-ORNL are being reannotated to improve the quality of their annotation available to the user community.

The JGI is made up of affiliates from a number of national laboratories including Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, Los Alamos National Laboratory, Oak Ridge National Laboratory, and the Hudson Alpha Genome Center.

genes in any given organism are still unknown. Our goal is to predict and experimentally verify the *in-vivo* function of proteins that lack homologs of known function ('unknown' protein families) and that are highly conserved between prokaryotes and plants. Our approach combines the extensive post-genomic resources of the plant field with the use of comparative genomic tools made possible by the availability of thousands of sequenced microbial genomes. This is an integrative approach to predict gene function whose early phase is computer-assisted, and whose later phases incorporate intellectual input from expert plant and microbial biochemists. It allows bridging of the gap between automated homology-driven annotations and the classical gene discovery efforts driven mainly by experimentalists. We have already analyzed 350 protein families and linked 50 of them to particular metabolic areas. In a second round of analysis we have predicted a testable function for 20 protein families and started experimental validation for 15 as summarized in Table 1. This has already led to experimentally supported annotations for >600 individual genes (Fig. 1A) and has the potential to yield such annotations for at least 3,000 more (Fig. 1B). From this work, which represents around half our planned analysis, it is clear that specific predictions are readily possible, and that these could never have been made without this type of integrative effort. It is also clear that this is a cost-effective way to assign function to unknown proteins. The predictions and status of the experimental validations will be available on a website in early 2009.

Fig. 1. Genes annotated so far (A) and potential for future annotations (B)



GTL

Phylogenomics-Guided Validation of Function for Conserved Unknown Genes

Valérie de Crécy-Lagard^{1*} (vcrecy@ufl.edu) and Andrew D. Hanson² (adha@ufl.edu)

¹Dept. of Microbiology and Cell Science and ²Dept. of Horticultural Sciences, University of Florida, Gainesville, Fla.

Identifying the function of every gene in all sequenced organisms is a central challenge of the post-genomic era. We are submerged in genomic, transcriptomic, and proteomic data but the functions of about half (range 20 to 60%) of the

Table 1. Progress in predicting and validating functions of unknown protein families

Unknown protein family (<i>Arabidopsis</i> AGI codes)	Predicted function	Experimental validation (genes tested)	Genes annotated Actual [Potential]
COG2154 (At1g29810, At5g51110)	PCD with accessory role in Moco metabolism	Completed ¹ (20)	233
COG0354, YgfZ (At4g12130, At1g60990)	Fe-S cluster synthesis/repair	In progress (7)	365
COG0720 (In Heterokonts, not higher plants)	Replaces FolB (& FolQ) in folate biosynthesis	In progress (9)	64
COG0212 (At1g76730)	Alternative 5-formyltetrahydrofolate cycloligase	In progress (5)	[46]
COG3404 (At2g20830)	Replaces 5-formyltetrahydrofolate cycloligase	In progress (6)	[66]
COG1836 (At1g78620)	Phytyl phosphate kinase	In progress (2)	[100]
COG3533 (At5g12960, At5g12950)	Glycosylhydrolase, plant cell wall breakdown	In progress (2)	[60]
COG3146 (At2g23390)	Pterin-dependent hydroxylase	In progress (1)	[200]
COG4319 (At3g09250, At4g10925)	Folate metabolism	In progress (1)	[81]
COG0009, YrdC (At5g60590)	t ⁶ A biosynthesis	In progress (5)	[703]
COG0533, YgiD (At2g45270, At4g22720)	t ⁶ A biosynthesis/telomere maintenance	In progress (1)	[730]
COG0009, YciO (At3g01920)	RNA metabolism	In progress (1)	[223]
COG0523 (At1g15730, At1g26520, At1g80480)	Metal chaperone, metal homeostasis	In progress (7)	[1135]
COG2016 At1g09150	acp ³ U in RNA	In progress (1)	[40]
COG2263 (At4g28830)	RNA methylase m ⁶ A in rRNA	In progress (1)	[40]
PROSC, YggS (At4g26860, At1g11930)	Proline biosynthesis	Pending	[622]
COG0624 (At5g43600)	Alternative N-formylglutamate deformylase	Pending	[26]
COG0697 (At3g02690)	Pterin efflux carrier	Pending	[36]
COG0451 (At1g19690)	Synthesis of sugars decorating lipid A	Pending	[436]
COG0277 (At4g36400)	Hydroxyacid dehydrogenase	Pending	[176]

* Presenting author

Comparative Genomics and Experimental Validation to Find Universal, Globally Missing Genes: The Universal Families COG009 and CO0533

Basma El Yacoubi^{1*} (basma@ufl.edu) and Valérie de Crécy-Lagard¹ (vcrecy@ufl.edu)

¹Dept. of Microbiology and Cell Science, University of Florida, Gainesville, Fla.

By combining comparative genomics-guided functional predictions based on plants and prokaryotes with experimental validations, we plan to predict and experimentally verify the function of 20 unknown protein families. This will permit annotation of >3000 individual genes. (See Valérie de Crécy-Lagard and Andrew D. Hanson presentations and poster by Anne Pribat). Here, we illustrate the approach with the case of two universal and globally missing genes. We exploited the availability of whole-genome sequence data, especially those of microorganisms with small genomes, to investigate the minimal gene set, and to find potentially globally missing genes for which biochemical or physiological data suggests universality.

t⁶A is a universal base modification occurring at position 37 in a subset of tRNAs decoding ANN codons. The biosynthesis pathway of this complex modification is not elucidated but is known to require threonine, ATP and bicarbonate (1,2). To date no gene has been associated with the t⁶A₃₇ biosynthesis pathway. Because it is universal and globally missing, traditional bioinformatic tools such as blast, phylogenetic occurrence and physical clustering cannot be applied to guide functional predictions. Instead, we generated a list of ortholog families present in a subset of the smallest genomes (total 16 genomes). This analysis generated a list of 95 orthologous families, only nine of these families did not have an experimentally verified function when this work was initiated and we focused on two of them for potential missing t⁶A genes for the reasons developed below.

1. Identification of COG0009 as a potential threonylcarbamoyladenine (t⁶A) enzyme

The COG0009 family was chosen as the first candidate for an involvement in t⁶A biosynthesis, because it has been shown to bind double-stranded RNA (3) and has been linked to translation phenotypes in both prokaryotes and eukaryotes (4,5), and also because of sequence homology with [NiFe] hydrogenase maturation protein HypF, which catalyzes a reaction with a chemistry similar to the one expected for a t⁶A enzyme (6). This gene family can further be split based on sequence comparison into three subfamilies: YrdC, Sua5 (YrdC with an extra domain termed Sua5) and YciO (Figure 2). One or two members of this family are present in each genome. The *A. thaliana* and *E. coli* genomes for example contain two, YrdC and YciO, while the yeast genome contains only one homolog, Sua5. We showed that 1) tRNAs from strains of *S. cerevisiae* lacking a *sua5* do not

contain t⁶A and that this phenotype is complemented by transforming with a plasmid encoding the wild type gene; 2) the homologs from *B. subtilis*, *M. maripaludis*, *E. coli yrdC*, but not the *E. coli yciO* are able to complement the t⁶A minus phenotype of the yeast Δ *sua5* and 3) the *yrdC* homolog is essential in *E. coli*, whereas *yciO* is not, and yeast lacking *sua5* are greatly impaired in their growth (information which was controversial in the literature); 4) *S. cerevisiae*, *B. subtilis*, *M. maripaludis* are able to complement the lethality phenotype of *yrdC* in *E. coli*, but not *E. coli yciO* and 5) *E. coli yrdC* is able to bind t⁶A apomodified tRNA^{Thr} but not unmodified transcript. Therefore, members of the YrdC/Sua5 family are most probably involved in t⁶A biosynthesis. It is unclear at this point which of the homologs of *A. thaliana* is a functional YrdC, and work is underway to investigate the plant family.



Figure 1. Domain organization of the YrdC/Sua5 family. Sua5 members have a well conserved YrdC domain but contain an addition domain, the Sua5 domain found only in the family of proteins. The YciO members are missing some conserved residues of the YrdC domain.

2. Identification of the COG0533 as another potential t⁶A enzyme

The biosynthesis of t⁶A requires multiple enzymatic steps and therefore, partners of YrdC should also be implicated in this pathway. These should follow the same phylogenetic distribution as COG0009. A candidate also identified through our bioinformatic analysis is COG0533. Interestingly, the domain conserved in COG0533 (YgjD domain) is also found in HypF and some HypF proteins contain both a YrdC and a YgjD domain. Also a YgjD domain is found in NodU and NodO which are carbamoyl transferases, chemistry that would resemble one of a t⁶A enzyme (Figure 2).

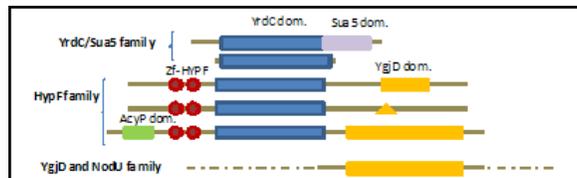


Figure 2. Conserved domain architecture of YrdC domain containing protein relevant to the t⁶A biosynthesis pathway. This was based on NCBI CDART and (1).

*Possible additional domains not shown.

To validate this family as a t⁶A biosynthesis enzyme, t⁶A content will be analyzed in yeast and *A. thaliana ygd* mutants and homologs from appropriately chosen Bacteria and Archaea will be tested for functional complementation. Also *in vitro* assays will be performed with recombinant

YrdC and YgjD to confirm their roles as partners in t⁶A biosynthesis. Finally the YgjD family in Eukaryotes contains two sub-families, the bacterial-like YgjD family (yeast YDL104C) and the archaeal and eukaryotic family (yeast YKR038C). The latter has been implicated in telomere length regulation (7). A member of each family from *Arabidopsis thaliana* and yeast will be tested for their potential role in t⁶A biosynthesis.

References

1. Aravind, L. and Koonin, E.V. (1999) Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *Journal of Molecular Biology*, 287, 1023-1040.
2. Elkins, B.N. and Keller, E.B. (1974) The enzymatic synthesis of N-(purin-6-ylcarbamoyl)threonine, an anticodon-adjacent base in transfer ribonucleic acid. *Biochemistry*, 13, 4622-4628.
3. Teplova, M., Tereshko, V., Sanishvili, R., Joachimiak, A., Bushueva, T., Anderson, W.F. and Egli, M. (2000) The structure of the *yrdC* gene product from *Escherichia coli* reveals a new fold and suggests a role in RNA binding. *Protein Sci.*, 9, 2557-2566.
4. Kaczanowska, M. and Ryden-Aulin, M. (2005) The YrdC protein: a putative ribosome maturation factor. *Biochim. Biophys. Acta*, 1727, 87-96.
5. Na, J.G., Pinto, I. and Hampsey, M. (1992) Isolation and characterization of *sua5*, a novel gene required for normal growth in *Saccharomyces cerevisiae*. *Genetics*, 131, 791-801.
6. Paschos, A., Bauer, A., Zimmermann, A., Zehlein, E. and Bock, A. (2002) HypF, a carbamoyl phosphate-converting enzyme involved in [NiFe] hydrogenase maturation. 277, 49945-49951.
7. Downey, M., Houlsworth, R., Maringe, L., Rollie, A., Brehme, M., Galicia, S., Guillard, S., Partington, M., Zubko, M.K., Krogan, N.J. *et al.* (2006) A genome-wide screen identifies the evolutionarily conserved KEOPS complex as a telomere regulator. *Cell*, 124, 1155-1168.

GTL

Genome Annotation: Coupling the Power of Plant-Prokaryote Comparative Genomics to Experimental Validation, COG0720 and COG3404

Anne Pribat^{1*} (pribat@ufl.edu), Linda Jeanguenin,¹ Valérie de Crécy-Lagard² (vcrecy@ufl.edu), and Andrew D. Hanson¹ (adha@ufl.edu)

¹Dept. of Horticultural Sciences and ²Dept. of Microbiology and Cell Science, University of Florida, Gainesville, Fla.

Project Goals: By combining comparative genomics-guided functional predictions based on plants and prokaryotes with experimental validations, we plan to predict and experimentally verify the function of 20 unknown protein families. This will permit annotation of >3000 individual genes.

The recent rapid expansion of genome sequencing projects has produced a mass of sequence data that are largely annotated automatically. With more than 800 genomes now completed and over 3,000 more in the pipeline, the propagation of unhelpful or incorrect annotations (e.g., 'hypothetical protein', generic annotations, annotation errors) is increasing exponentially. This degrades databases and so prevents optimal use of genome information.

By combining comparative genomics-guided functional predictions based on plants and prokaryotes with experimental validations, we plan to predict and experimentally verify the function of 20 unknown protein families. This will permit annotation of >3000 individual genes.

Here, we illustrate the approach with unknown protein families COG0720 and COG3404.

1. COG0720: A novel enzyme in folate biosynthesis

Certain lower plants (diatoms, pelagophytes) and many bacteria lack the folate synthesis enzyme dihydroneopterin aldolase (FolB), which converts dihydroneopterin to 6-hydroxymethyl-dihydropterin (HMDHP) in the folate biosynthesis pathway (Fig. 1). Comparative genomic analysis showed that most plants and bacteria that lack FolB have a paralog (COG0720) of the tetrahydrobiopterin synthesis enzyme 6-pyruvoyltetrahydropterin synthase (PTPS) in which a glutamate replaces or accompanies the canonical catalytic cysteine residue. A similar COG0720 protein from the malaria parasite *Plasmodium falciparum* (which is related to plants) was recently shown to form HMDHP from dihydroneopterin triphosphate *in vitro*, and was proposed to provide a bypass to the FolB step *in vivo* (Fig. 1) (Dittrich *et al. Mol. Microbiol.* **67**: 609).

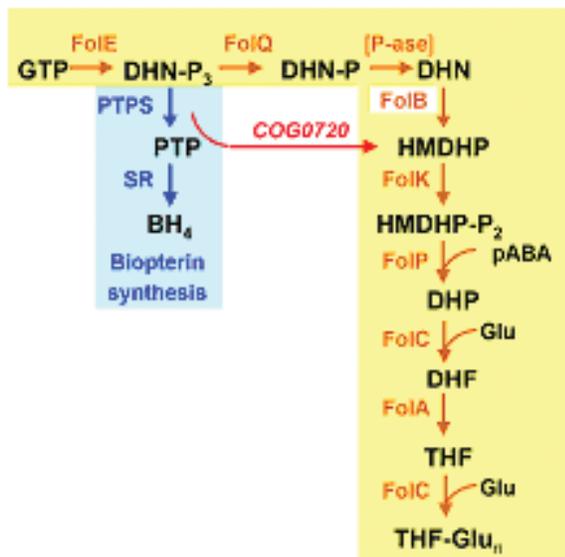


Figure 1. Folate synthesis pathway

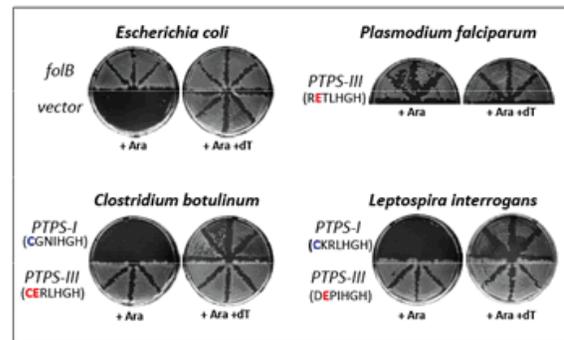


Figure 2. *E. coli* ΔfolB complementation by PTPS-III proteins

Bacterial COG0720 genes encoding proteins with active-site glutamate or glutamate plus cysteine (PTPS-III proteins), or cysteine alone (PTPS-I proteins), were accordingly tested along with the *P. falciparum* sequence for complementation of the *Escherichia coli* folB mutation. The *P. falciparum* sequence and other PTPS-III proteins were active; PTPS-I proteins were not (Fig. 2). These results establish that COG0720 proteins indeed functionally replace FolB.

2. COG3404: A probable novel folate metabolism enzyme

Folates are essential as cofactors for one-carbon (C₁) transfer reactions in nucleotide biosynthesis, amino acid metabolism, and elsewhere. 5-Formyltetrahydrofolate (5-CHO-THF), formed by a side-reaction of serine hydroxymethyltransferase (SHMT), is not a C₁ donor but a potent inhibitor of various folate-dependent enzymes. It must consequently be metabolized. The only enzyme known to do this is 5-CHO-THF cycloligase (5-FCL, Fig. 3). However, there are almost surely alternative enzymes that metabolize 5-CHO-THF because knocking out 5-FCL in plants and microorganisms gives rather mild phenotypes (Goyer *et al., J. Biol. Chem.* **280**: 26137; Holmes *et al., J. Biol. Chem.* **277**: 20205). Comparative genomics analysis identified a strong candidate: COG3404, a formiminotransferase-like protein, which potentially mediates transfer of the formyl group of 5-CHO-THF to glutamate. This prediction can be validated by testing whether COG3404 substitutes for 5-FCL in complementation assays using an *E. coli* 5-FCL (*ygfA*) deletant.

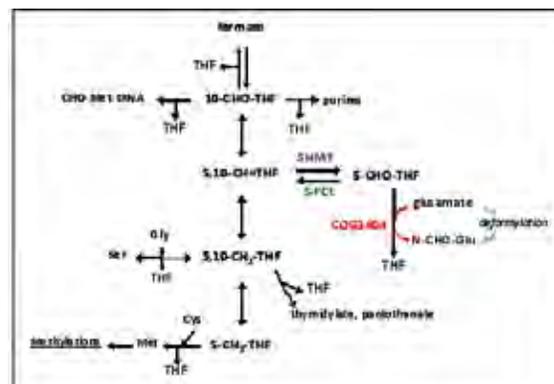


Figure 3. Folate metabolism overview

Use of Modern Chemical Protein Synthesis Techniques to Experimentally Validate the Functional Annotation of Microbial Genomes

Stephen Kent* (skent@uchicago.edu) and Youhei Sohma

Institute for Biophysical Dynamics, University of Chicago, Chicago, Ill.

Project Goals: To develop high throughput chemical methods to make large numbers of predicted proteins and protein domains, based on microbial genome sequences.

Chemical protein synthesis is a powerful way of studying the properties of predicted proteins. It involves the use of organic chemistry to construct a predicted polypeptide chain from protected amino acid starting materials, followed by folding of the synthetic polypeptide to give the unique, defined tertiary structure of the protein molecule. The synthetic protein is then used to experimentally validate the predicted biochemical function, and in selected cases to determine the X-ray structure of the protein molecule (Figure 1).

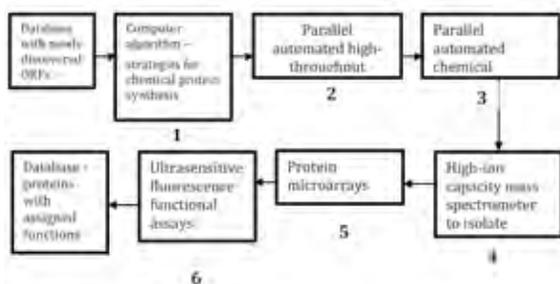
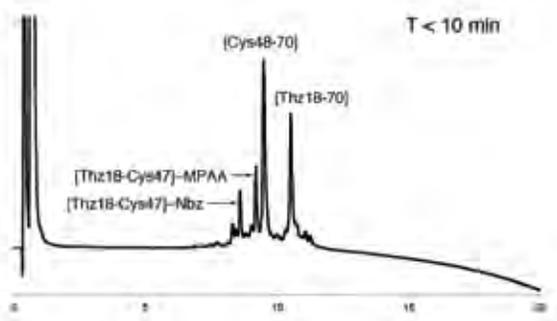
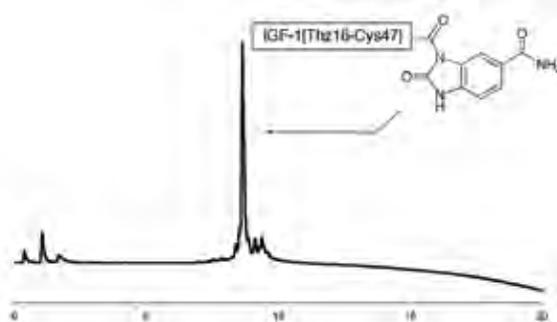
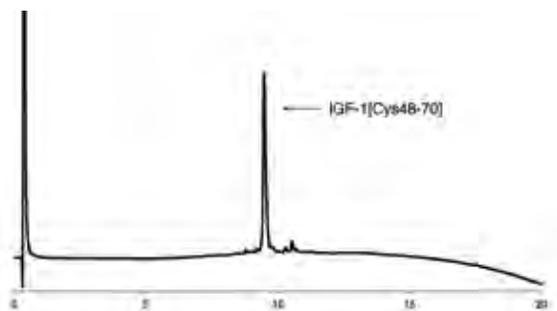


Figure 1. Modular high-throughput platform for fast and parallel total chemical synthesis, mass-spectrometric purification and single-molecule spectroscopic assay to annotate function for newly predicted proteins.

Modern total protein synthesis has evolved from the 'chemical ligation' methods introduced by the Kent laboratory in the mid-1990s [Dawson PE, Kent SB: Synthesis of native proteins by chemical ligation. *Annual Review of Biochemistry* 2000, **69**:923-960.]. Unprotected synthetic peptide segments, spanning the amino acid sequence of the target polypeptide chain, are covalently joined to one another by chemo-selective reaction of unique, mutually reactive functional groups on each segment. Native chemical ligation, the thioester-mediated covalent bond-forming chemoselective reaction of unprotected peptides at a Cys residue, is the most robust and useful ligation chemistry developed to date.

Synthesis of peptide-thioesters. For many biomedical researchers the utility of chemical protein synthesis based on native chemical ligation methods is limited by the inability to make peptide-thioesters. Routine synthesis of peptide-thioesters has not previously been possible using Fmoc chemistry

SPPS methods [Camarero JA, Mitchell AR: Synthesis of proteins by native chemical ligation using Fmoc-based chemistry. *Protein and Peptide Letters* 2005, **12**:723-728]. In the first stage of prototyping high throughput chemical protein synthesis, we have used x,y,z robotics and laboratory automation to develop efficient Fmoc chemistry SPPS protocols for the simultaneous parallel synthesis of the key peptide-thioester building blocks needed for chemical protein synthesis. This made use of a recently reported novel resin linker [Blanco-Canosa JB, Dawson PE: An efficient Fmoc-SPPS approach for the generation of thioester peptide precursors for use in native chemical ligation. *Angew Chem Int Ed Engl.* 2008, **47**:6851-5]. Typical data are shown in Figure 2 (Top).



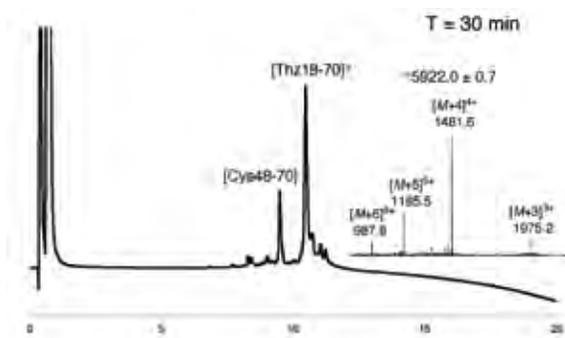


Figure 2. (Top): automated robotic Fmoc SPPS preparation of peptide-thioesters. HPLC-electrospray MS of crude products are shown. (Bottom): LCMS analyses of native chemical ligation of the two peptides; near-quantitative reaction after 30 minutes was observed. The 53 amino acid residue ligated product had a mass of 5922 Daltons.

We expect that this approach to parallel peptide synthesis will be satisfactory for the simultaneous production of ~100 peptide-thioester segments, each containing 30 to 40 amino acid residues. Ready preparation of peptide-thioesters enables the straightforward total chemical synthesis of proteins by native chemical ligation **Figure 2 (Bottom)**. Proof-of-concept total chemical syntheses of several predicted proteins from microbial genomes will be presented.

GTL

Towards Annotation of the Unannotated—A Dissection of Unannotated Proteins in *Clostridium thermocellum*

Y.J. Chang,* Loren J. Hauser, Gwo-liang Chen, Frank Larimer, Pavan Ghatty, and **Miriam Land** (ml3@ornl.gov)

Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

Project Goals: In all prokaryotic genomes sequenced there is a substantial fraction of genes that are uncharacterized or unknown, which poses one of the biggest hurdles to the development of genome annotation, thus genome science. The annotation of unknown proteins remains an unsolved problem.

Attempts to solve this problem have been limited, due to the lack of methodology, also the overwhelming interest in immediate opportunity of analyzing known sequences. In this study, we systematically explored the unannotated genes in genome of *Clostridium thermocellum* ATCC 27405. The approaches we applied include genomic context, gene order comparison, protein structure and phylogenetic analysis. With the help of extensive searches across all other genomes and horizontal gene transfer (HGT)/transposable elements (TE) /repeat detection, clusters of functionally

related unknown genes are identified. A possible relationship between unknown and other neighboring genes, such as HGT, protein splicing site, transposable elements is discussed.

This whole genome initiative aimed at 1) initiating annotation of the unannotated at genome scale 2) mapping out the function and possibly regulatory relationships of unknown proteins, and 3) eventually defining the uncharacterized proteins. The work presented here represents a part of the ongoing efforts to annotate the unannotated at the Oak Ridge National Laboratory.

In all prokaryotic genomes sequenced there is a substantial fraction of genes (up to 60%) that are uncharacterized or unknown, which poses one of the biggest hurdles to the development of genome annotation, thus genome science. The annotation of unknown proteins remains an unsolved problem.

Attempts to solve this problem have been limited, due to the lack of methodology, the overwhelming interests in immediate opportunity of analyzing known proteins. In this study, we systematically explored the unannotated genes in genome of *Clostridium thermocellum* ATCC 27405. The approaches applied include genomic context, gene order comparison, protein structure and phylogenetic analysis. With the help of extensive searches across all other genomes and horizontal gene transfer (HGT)/repeat detection, clusters of functionally related unknown genes are identified. A possible relationship between unknown and other neighboring genes, such as HGT, protein splicing site, transposable elements is discussed.

This whole genome initiative aimed at 1) initiating annotation of the unannotated at genome scale 2) mapping out the function and possibly regulatory relationships of unknown proteins, and 3) eventually defining the uncharacterized proteins. The work presented here represents a part of the ongoing efforts to annotate the unannotated at the Oak Ridge National Laboratory.

GTL

An Integrated Approach to Experimental Validation of Putative Gene Functions in *M. acetivorans*

Ethel Apolinario,¹ Yihong Chen,² Zvi Kelman,² Zhuo Li,² Basil J. Nikolau,³ Kevin Sowers,¹ Eric Testroet,³ and **John Orban**^{2,*} (orban@umbi.umd.edu)

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

Project Goals: The goal of the project is to develop an integrated high-throughput approach to functionally

annotate a large group of conserved hypothetical genes in the methanogenic archaeon, *Methanosarcina acetivorans*. The focus will be 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.

We are developing an integrated approach to facilitate experimental determination or validation of enzymatic functions in the methanogenic archaeon, *Methanosarcina acetivorans* (MA). The goal is to use a combination of gene knockouts, metabolic profiling with mass spectrometry, NMR-based ligand screening, and biochemical assays to accurately annotate gene targets.

Before choosing suitable initial targets for experimental validation, we first manually re-annotated ~710 of the 4524 MA genes based on updated annotations in other Archaea and a thorough search for experimental data in the literature. This process led to ~220 gene annotations which were more specific than the original function assignment, ~60 gene annotations became less specific, and ~430 gene annotations were not altered. The set of genes with a more specific functional annotation was further surveyed for 1) putative enzymatic function, 2) *in vivo* expression in MA from existing proteomic and DNA microarray data, and 3) transmembrane containing regions. Approximately 30 of these gene products were putatively identified as soluble enzymes with detectable *in vivo* expression and were targeted for cloning, expression screening and purification. Genes that could be expressed at reasonable levels of soluble protein in *E. coli* were further targeted for preparation of knockout alleles in MA. Alternatively, if an *E. coli* homolog exists, a non-essential gene knockout will be obtained directly from the Keio collection.

Expressed proteins are screened for putative substrates and products using a one-dimensional ¹H NMR-based assay, waterLOGSY (water-Ligand Observed via Gradient Spectroscopy), which has been used in the past for drug screening applications (1). Additionally, we are using mass spectrometry to profile metabolite differences in *E. coli* strains containing the relevant MA gene and we will profile metabolites in MA gene knockout strains. Genetic complementation of *E. coli* knockout strains is also being used to confirm the function of MA genes. Results on genes putatively identified as being involved in biotin metabolism and regulation will be highlighted.

Reference

1. Dalvit, C., Fogliatto, G., Stewart, A., Veronesi, M., and Stockman, B. (2001) WaterLOGSY as a method for primary NMR screening: practical aspects and range of applicability, *J Biomol NMR* 21, 349-359.

Transcriptome Analysis of *Chlamydomonas reinhardtii* using Ultra-High-Throughput Sequencing

David Casero Díaz-Cano,¹ Madeli Castruita,² Sabeeha Merchant,² and Matteo Pellegrini^{1*} (matteop@mcdb.ucla.edu)

¹Dept. of Molecular, Cell and Developmental Biology and ²Dept. of Chemistry and Biochemistry, University of California, Los Angeles, Calif.

Project Goals: The advent of massively parallel short read sequencing technology opens the door to (near) full coverage of the *Chlamydomonas* transcript map via deep sequencing of mRNAs. To evaluate the potential of Illumina's Solexa technology for a) generating a whole transcriptome for *Chlamydomonas*, b) identifying differentially expressed genes, and c) reconstructing gene models de novo, we analyzed RNAs isolated from metal deficient conditions and developed novel algorithms for data analysis.

Chlamydomonas reinhardtii, a unicellular eukaryote in the plant lineage, has been exploited in the laboratory over the last 50 years as a model organism for the study of eukaryotic photosynthesis. Unlike flowering plants, *Chlamydomonas* synthesizes and maintains a functional photosynthetic apparatus even when grown in the dark by respiration on organic carbon. This means that mutants with defects in either the light or dark reactions of photosynthetic metabolism can be maintained and characterized biochemically. Accordingly, many of the fundamental discoveries leading to today's knowledge of photosynthesis are derived from the application of biochemical and classical genetic approaches using the *Chlamydomonas* model. In the last decade, pathways of energy metabolism beyond photosynthesis have received considerable attention from the research community, specifically the biosynthesis of H₂ and the production of ethanol and other fermentation products resulting ultimately from solar energy conversion in the photosystems.

The genome of *Chlamydomonas* consists of 121 Mb in 17 chromosomes. Relative to other eukaryotes, a typical *Chlamydomonas* gene is intron-rich; there are 8.3 exons per gene and the average intron size is 373 bp. These characteristics make de novo prediction of gene models very difficult in the absence of a high quality dense transcript map. The existing datasets cover only 8631 (about half) of the 15,143 predicted protein-coding gene models, and only half of these include full-length coverage. Accordingly, despite the importance of *Chlamydomonas* as a model for the study of photosynthesis and energy metabolism, **only a quarter of the protein-coding gene models are accurately computed and verified via a transcript map.**

The advent of massively parallel short read sequencing technology opens the door to (near) full coverage of the *Chlamydomonas* transcript map via deep sequencing of mRNAs. To

evaluate the potential of Illumina's Solexa technology for a) generating a whole transcriptome for *Chlamydomonas*, b) identifying differentially expressed genes, and c) reconstructing gene models *de novo*, we analyzed RNAs isolated from metal deficient conditions

We have verified that these libraries may be used to quantitatively estimate transcript fold changes in different conditions using existing gene models. We are also developing a new annotation pipeline using only the short read sequencing data, and have shown that even simple approaches allow us to accurately reconstruct a set of manually curated genes. This approach therefore promises to not only measure transcript counts and differential expression but also comprehensively annotate the genomes of organisms for which we have only partial genome sequences.

—
GTL

Transcript Verification Coupled with Metabolic Network Modeling for *Chlamydomonas reinhardtii*

Ani Manichaikul,^{1*} Lila Ghamsari,^{2,3} Chenwei Lin,^{2,3} Erik F.Y. Hom,⁴ Ryan R. Murray,^{2,3} Roger L. Chang,⁵ Arvind K. Chavali,¹ Yun Shen,^{2,3} Xinping Yang,^{2,3} Ines Thiele,⁵ Jason A. Papin¹ (papin@virginia.edu), and **Kourosh Salehi-Ashtiani**^{2,3} (Kourosh_Salehi-ashtiani@DFCI.harvard.edu)

¹Dept. of Biomedical Engineering, University of Virginia Health System, Charlottesville, Va.; ²Center for Cancer Systems Biology (CCSB), Dept. of Cancer Biology, Dana-Farber Cancer Institute, Boston, Mass.; ³Dept. of Genetics, Harvard Medical School, Boston, Mass.; ⁴Dept. of Molecular and Cellular Biology, Harvard University, Cambridge, Mass.; ⁵Dept. of Bioengineering, University of California San Diego, La Jolla, Calif.

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

***reinhardtii* by RT-PCR and RACE, functionally validate the transcripts by yeast two-hybrid experiments, and build and interpret predictive metabolic network models based on the obtained results.**

With genome sequencing of many bioenergy organisms completed or in progress, there is a growing need to bridge the gaps between primary sequence information, gene annotation, and mapping of metabolic networks. The release of *C. reinhardtii* genome sequence has made this unicellular algae a viable target for metabolic engineering towards improved biofuel production; however, the relevant genes are not validated and the needed metabolic network map is not currently available. We present an update on our integrative process of coupling network reconstruction with transcript verification on *C. reinhardtii*. Using literature (and other resources) we have reconstructed a model of the central metabolic network of *C. reinhardtii*. Our *in silico* reconstructed network encompasses 252 reactions, 109 metabolites, and 295 gene products, and accounts for their intracellular compartmentalizations. Through systematic comparison of *JGI* transcript annotation against publicly available protein sequence databases, we identified and assigned E.C. numbers to transcript sequences of all but 12 of the hypothesized gene-associated reactions. To validate the involved gene products, we have carried out open reading frame (ORF) verification by RT-PCR and RACE (rapid amplification of cDNA ends) for all the protein coding genes involved (as well as a set of positive control ORFs). Following optimization of the RT-PCR procedure for high GC content of *C. reinhardtii* transcriptome, we were able to verify (by cloning and sequencing) approximately 80% of the ORFs annotated in the central metabolic map. Data from RACE experiments is being used to further verify and refine annotation of the transcripts. The generated metabolic model, now supported by experimental results, carries *in silico* 'growth' with a yield of 0.012 g DW / mmol acetate, consistent with the experimentally derived value for the organism. The network, along with ORF verification experiments is now being expanded to include all major metabolic reactions, including (but not limited to), fatty acid, isoprenoid, carotenoid and other hydrocarbon pathways essential to certain biofuel production. The metabolic ORF clone resource that we have generated will be made available without restrictions to the research community.

Genemap-MS: High Throughput Mass Spectrometry Approaches to Microbial Gene Annotation Validation

Trent R. Northen* (trnorthen@lbl.gov), Benjamin Bowen, Steven M. Yannone, Jill Fuss, John Tainer, and **Gary Siuzdak** (siuzdak@scripps.edu)

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

Project Goals: 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 our expertise with developing and applying MS technologies to generate datasets directly applicable to validating currently annotated genomes. The technologies represent multifaceted and universally applicable approaches to validate: 1) protein expression 2) metabolic pathways and 3) biochemical activities. The datasets from these analyses are integrated into computational metabolic networks to provide functional validation of the many hypothetical 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 microbes for bioenergy production.

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. Project Genemap-MS is developing and applying 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 (NIMS and Nimzyme). To maximally improve existing homology based annotations, we are using diverse model systems which span the three branches of the tree of life *Synechococcus sp* (cyanobacteria), *Sulfolobus solfataricus* (archaea), *Cblamydomonas Reinhardtii* (eukaryota); an integrated metabolomics/proteomics pipeline; and our new Meta-IQ bioinformatics software.

Novel Mass Spectrometry Based Platforms for the Investigation of Model Organisms

Sunia A. Trauger^{1*} (strauger@scripps.edu), Trent Northen,² Nitin Baliga,³ Steven M. Yannone,² Michael W.W. Adams,³ and **Gary Siuzdak**¹

¹Dept. of Molecular Biology and Center for Mass Spectrometry, The Scripps Research Institute, La Jolla, Calif.; ²Dept. of Molecular Biology, Lawrence Berkeley National Laboratory, Berkeley, Calif.; and ³Institute for Systems Biology; University of Georgia, Athens, Ga.

Project Goals: The objective of the MAGGIE Program is to comprehensively characterize the Protein Complexes (PCs) and Modified Proteins (MPs) underlying microbial cell biology. MAGGIE will address immediate GTL missions by accomplishing three specific goals: (1) to provide a comprehensive, hierarchical map of prototypical microbial PCs and MPs by combining native biomass and tagged protein characterizations from hyperthermophiles with comprehensive systems biology characterizations of a non-thermophilic model organism, (2) to develop and apply advanced mass spectroscopy and SAXS technologies for high-throughput characterizations of PCs and MPs, (3) to create and test powerful computational descriptions for protein functional interactions.

The Scripps Center for Mass Spectrometry is focused on developing methodologies for the characterization of protein-protein complexes and the elucidation of functional pathways affected by specific perturbations for the three model organisms of interest: *Pyrococcus furiosus*, *Sulfolobus solfataricus* and *Halobacterium salinarum*. The different mass spectrometry based platforms we have developed include, capillary LC based metabolomics, high-throughput proteomics approaches as well as nanostructure initiator mass spectrometry (NIMS) which is a new mass spectrometry based approach which shows considerable promise for the monitoring of a specific enzymatic activity in microbial communities.

The field of global mass-based metabolomics provides a platform for discovering unknown metabolites and their specific biochemical pathways. We report the identification of a new endogenous metabolite, N(4)-(N-acetylaminopropyl)spermidine which was found to be up-regulated and the use of a novel proteomics based method for the investigation of its protein interaction using metabolite immobilization on agarose beads. The metabolite was isolated from the organism *Pyrococcus furiosus*, and structurally characterized through an iterative process of synthesizing candidate molecules and comparative analysis using accurate mass LC-MS/MS. An approach developed for the selective preparation of N(1)-acetylthermospermine, one of the possible structures of the unknown metabolite, provides a convenient route to new polyamine derivatives through methylation on the

N(8) and N(4) of the thermospermine scaffold. The biochemical role of the novel metabolite as well as that of two other polyamines: spermidine and agmatine is investigated through metabolite immobilization and incubation with native proteins. The identification of eleven proteins that uniquely bind with N(4)-(N-acetylaminopropyl)spermidine, provides information on the role of this novel metabolite in the native organism. Identified proteins included hypothetical ones such as PF0607 and PF1199, and those involved in translation, DNA synthesis and the urea cycle like translation initiation factor IF-2, 50S ribosomal protein L14e, DNA-directed RNA polymerase, and ornithine carbamoyltransferase. The immobilization approach demonstrated here has the potential for application to other newly discovered endogenous metabolites found through untargeted metabolomics, as a preliminary screen for generating a list of proteins that could be further investigated for specific activity.

We have also developed a Nanostructure-Initiator Mass Spectrometry (NIMS) enzymatic (Nimzyme) assay in which enzyme substrates are immobilized on the mass spectrometry surface by using fluororous-phase interactions. This “soft” immobilization allows efficient desorption/ionization while also enabling the use of surface-washing steps to reduce signal suppression from complex biological samples, which results from the preferential retention of the tagged products and reactants. The Nimzyme assay is sensitive to subpicogram levels of enzyme, detects both addition and cleavage reactions (sialyltransferase and galactosidase), is applicable over a wide range of pHs and temperatures, and can measure activity directly from crude cell lysates. The ability of the Nimzyme assay to analyze complex mixtures is illustrated by identifying and directly characterizing β -1,4-galactosidase activity from a thermophilic microbial community lysate.

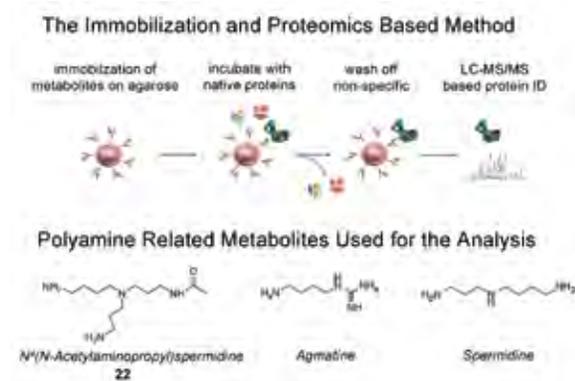


Figure 1. The metabolite immobilization, incubation with cell lysate, followed by proteomic analysis allows a first look at identifying proteins which interact with a novel metabolite of interest.

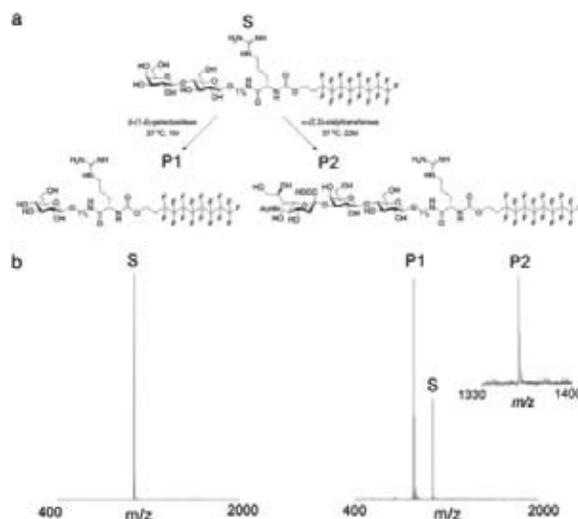


Figure 2. On-chip NIMS enzymatic activity assay (Nimzyme assay). (a) Substrate (b) Mass spectra of the substrate (Left) and resulting products (Right).

GTL

Annotation of Translation Initiation Sites Using Prodigal

Doug Hyatt, **Miriam Land** (ml3@ornl.gov), and Loren Hauser*

Oak Ridge National Laboratory, Oak Ridge, Tenn.

<http://prodigal.ornl.gov/>

Project Goals: Provide an update on the microbial gene finding effort at Oak Ridge.

Last year, ORNL introduced the microbial genefinding program Prodigal (Prokaryotic Dynamic Programming Genefinding Algorithm). Since that time, Prodigal has been incorporated into the Joint Genome Institute annotation pipeline. Prodigal has been used to annotate all microbial organisms submitted to Genbank by JGI in 2008, resulting in an enormous amount of data with which to measure the accuracy of the algorithm and to make necessary improvements. Prodigal is consistently being improved as new discoveries are made and more data is collected, and new versions are released every couple of months containing the updated changes to the program.

In the course of reviewing the pipeline annotations, we discovered that while Prodigal did quite well on locating translation initiation sites in 85-90% of the genomes, it experienced difficulties in those genomes that do not use the canonical Shine-Dalgarno ribosomal binding site motif (AGGAGG), or some other closely related motif. An effort was made to identify and classify the genomes which do not use this motif (or at least do not use it often). We consulted the literature on translation initiation sites, as well as exam-

ining the full set of finished microbial genomes in Genbank computationally to look for novel motifs. A new version of Prodigal was constructed with a much more complex RBS motif-finding system able to discover novel motifs while not abandoning its knowledge of the default Shine-Dalgarno motif (as happens in many other genefinding programs which auto-discover motifs). This version has since been incorporated into the JGI annotation pipeline.

In consulting the literature, we found that, in *Crenarchaea*, the first gene in an operon often has no ribosomal binding site motif, but genes internal to an operon often do use an SD motif. In *Aeropyrum pernix*, and some other archaea, a GGTG motif was observed computationally. This motif was strong and present in well over 50% of the start sites. Many chlorobi and cyanobacteria were observed to use the SD motif extremely infrequently, and minor AAAA/TATA type motifs were often found 13-15bp upstream (which may be transcriptional in nature). In one organism in particular, the bioenergy-related *Flavobacterium johnsoniae*, literature had documented a strong TAAA motif close to the start codon. This finding was confirmed by our computational analysis. The challenge, after identifying these patterns, was to create a flexible translation initiation site evaluator capable of auto-discovering these novel motifs while not losing sight of the still occasionally used Shine-Dalgarno motif (which could be used so infrequently that it would never be found with motif finding, but is still present in a significant percentage of genes, such as 2-3%).

We approached the problem by creating a motif finder that auto-discovered motifs in the RBS region of length 3-6bp,

with the restriction that all 3bp subsets of that motif had to be present in at least 20% of the genes. However, we did allow one mismatch in 5bp and 6bp motifs. The program used an iterative algorithm similar to Prodigal's default Shine-Dalgarno algorithm to assign log-likelihood weights to each motif. We then took this motif finder and Prodigal's default Shine-Dalgarno motif finder and combined them into a single TIS scorer with three distinct cases. In the first instance, if the organism used the SD motif frequently, we used Prodigal's existing default SD scorer. In the second instance, if a novel strong motif, such as GGTG in *Aeropyrum pernix*, was discovered, we used the new scoring system. Finally, if no strong motif of any kind was found, but some weak motifs were found, we used both scoring systems and took the maximum result. In *Crenarchaea*, for example, non-SD motif genes at the beginning of operons will get a decent score from the new scoring system, whereas the internal SD-motif-using genes in operons will get a good score from the old scoring system. This new version of Prodigal was tested on *Cyanobacteria*, *Chlorobi*, *Crenarchaea*, and GGTG-using *Euryarchaea*, and found to outperform the previous version of Prodigal and a version created that only used the new scoring system (but not the old one). The final version of the new TIS finder was completed, and the new version of Prodigal was introduced into the JGI pipeline in December, 2008.

Prodigal is routinely run on all finished genomes in Genbank every couple of months, and detailed comparisons with the Genbank files are performed. This data is available at the Prodigal website (<http://prodigal.ornl.gov/>).

Computing Resources and Databases

—
GTL

Release of Taxomatic and Refinement of the SOSCC Algorithm

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

¹Michigan State University, East Lansing, Mich.;
²American Type Culture Collection, Manassas, Va.; and
³University of Michigan, Ann Arbor, Mich.

Project Goals: The Taxomatic was conceived as a tool for aiding in the interpretation of large-scale phylogenetic trees, and uncovering unresolved nomenclatural and placement errors in large sets of sequence data using visualization techniques drawn from the field of exploratory data analysis. In this phase of the work we have taken what we have learned from early prototypes and deployed the tool as a web based service available through the RDP.

The Taxomatic was conceived as a tool for aiding in the interpretation of large-scale phylogenetic trees, and uncovering unresolved nomenclatural and placement errors in large sets of sequence data using visualization techniques drawn from the field of exploratory data analysis. The tool allows a user to generate a heatmap image (see Figure 1) representing the pair-wise similarity between large numbers of rRNA sequences (or any other set of homologous genes or concatenated genes) in the context of an existing taxonomy. The Self-Organizing Self-Correcting Classifier (SOSCC) is an algorithm that was developed as an extension to the Taxomatic to programmatically reorder similarity matrices and, based on the reordered matrix, identify and resolve inconsistencies within a proposed taxonomy. Together, these two tools allow a user to: 1) rapidly identify taxonomic anomalies (misplaced sequences), and 2) correct these anomalies.

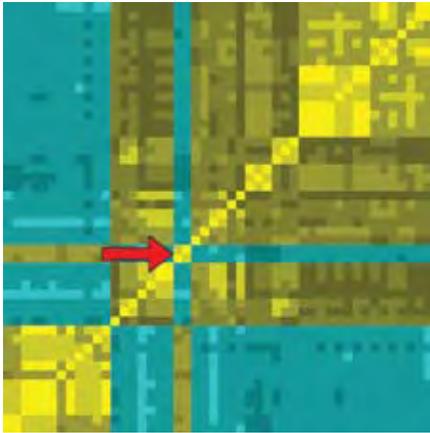


Figure 1. Heatmap of the current taxonomy for the families *Rhodocyclaceae* (Betaproteobacteria) and *Brucellaceae* (Alphaproteobacteria). A taxonomic anomaly (the genus *Shinella*) can be seen near the center of the image identifiable by the lines of cyan inside the otherwise yellow *Rhodocyclaceae* group.

For input data, the Taxomatic (<http://rdp.cme.msu.edu>) can use internally retrieved sets of RDP or *my*RDP sequences, externally generated FASTA sequence alignment files, or DNADist files. Within the Taxomatic interface, mouse-over action by a user will trigger the graphical outlining of a taxonomic overlay to highlight specific groups in the input taxonomic hierarchy. The taxonomic hierarchy can either be provided to the user as the default RDP taxonomy or can be directly supplied by the user in XML format. As a third alternative, for the purpose of an unlabelled overview, users may choose to proceed without a selected taxonomy. For higher zoom levels on the heatmap of the Taxomatic interface, detailed strain information is made available to the user through pop-up windows. Users can save the full set of input conditions necessary to either reproduce or extend their current heatmap analysis including: a copy of their heatmap (image size depends on the current zoom level), the matrix used to generate the heatmap, and the taxonomic hierarchy. The taxonomic hierarchy is used as both an algorithmic aid for ordering the sequences in the heatmap, and as a means for generating coordinates on the interface for plotting group overlays.

The SOSCC implements a two-step process in which the source matrix is first reorganized, after which the reordered sequences are reclassified. The source matrix is reorganized in an iterative process that results in an output matrix where closely related sequences are placed adjacent to each other, based on sequence similarity rather than presumptive taxonomic identity. The second phase can proceed along two different approaches. The first approach adjusts the borders of the source taxonomy on the output matrix based on how many adjacent sequences were originally members of the same group. The second approach identifies groups in the output matrix in an unsupervised fashion by identifying large changes in sequence distances. The SOSCC can accept any source data that is acceptable to the Taxomatic and, by default, displays the resulting matrix through the Taxomatic user interface. One option allows for running the

SOSCC on 100 bootstrapped samplings drawn from the set of aligned input sequences. With the bootstrapped sampling option, only those taxonomic rearrangements supported by a user-defined threshold level of percentage consistency with the rearrangement will be retained. Bootstrapping is currently limited to 2000 sequences, and can only be used when both the sequence data and taxonomic hierarchy are supplied. When using the bootstrapping feature of the Taxomatic, results are provided to the user via e-mail. These results include: a detailed report on the classification of each sequence; links to view the original and resulting matrix on Taxomatic; and a link to download the resulting matrix.

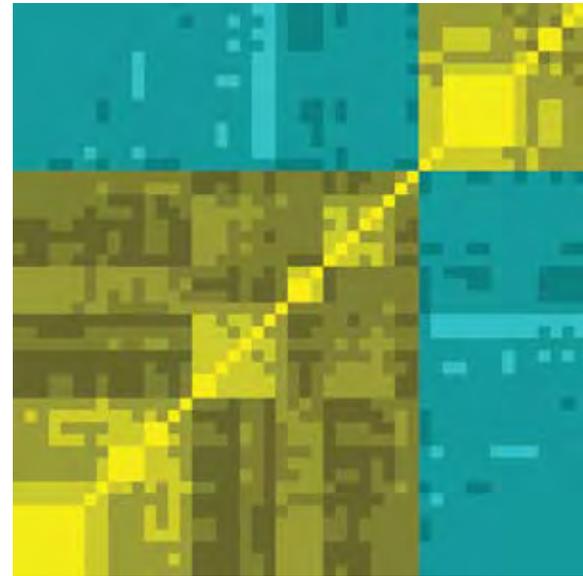


Figure 2. Distance matrix from Fig. 1 preprocessed with SOSCC. *Rhodocyclaceae* (lower left) and *Brucellaceae* (upper right). The discontinuity in the image is gone as *Shinella* has moved to the upper right corner of the image.

The Taxomatic and the SOSCC are both implemented as java servlets, exposing the functionality of their core classes to users using the Spring 2.5 MVC framework. The heatmap image that is displayed by the Taxomatic is delivered as a collection of tiles generated on the fly by the server. The taxonomic hierarchy information is loaded from the server for the currently displayed heatmap. Sequence information is fetched from the NamesforLife (<http://names4life.com>) database as the user mouse-overs individual sequences at $\geq 8x$ zoom levels. All of the presentation code is written using a combination of JSTL and javascript.

Both the Taxomatic and SOSCC have SOAP Web Services ports written in java using the JAX-WS reference implementation. The Taxomatic web service exposes methods to submit a similarity matrix or aligned sequence data to be displayed on the Taxomatic. When aligned sequence data are supplied, an uncorrected distance matrix is calculated from the sequences. The SOSCC web service exposes methods that submit aligned sequence data or a similarity matrix for processing based on either a default set of options or caller-specified set of options.

References

1. J. R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* (Advance Access November 2008). DOI 10.1093/nar/gkn879.
2. G. M. Garrity, and T. G. Lilburn. Self-organizing and self-correcting classifications of biological data. *Bioinformatics* (2005) 21:2309–2314.

This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy under Contract No. DE-FG02-04ER63933

SBIR

NamesforLife Semantic Resolution Services for the Life Sciences

Charles T. Parker,¹ Sarah Wigley,¹ Nicole Osier,² Jordan Fish,¹ Qiong Wang,¹ Donna McGarrel,¹ James R. Cole,^{1,2} Catherine Lyons,¹ and **George M. Garrity**^{1,2*} (garrity@names4life.com)

NamesforLife, LLC, and Michigan State University, East Lansing, Mich.

Project Goals: NamesforLife is a novel technology that resolves uncertainty about the meaning of biological names or other dynamic terminologies. It uses those terms to create persistent links to related information, goods, and services available on the Internet, even if the terms have changed.

Within the Genomes-to-Life Roadmap, the DOE recognizes that a significant barrier to effective communication in the life sciences is a lack of standardized semantics that accurately describe data objects and persistently express knowledge change over time. As research methods and biological concepts evolve, certainty about correct interpretation of prior data and published results decreases because both become overloaded with synonymous (multiple terms for a single concept) and polysemous terms (single terms with multiple meanings). Ambiguity 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 consists of an ontology, an XML schema, and an expertly managed vocabulary coupled with Digital Object Identifiers (DOIs) to form a semantic resolution service that disambiguates terminologies, presents them to end-users in a temporal context and persistently links them to relevant resources and services on the Internet. Our initial implementation of N4L technology is for the validly published biological 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 in a given situation.

Biological names also serve as key terms used to index and access information in databases and the scientific, technical, medical, and regulatory literature. Understanding the correct meaning of a biological name, in the appropriate context, is essential. This is not a trivial task, and the number of individuals with expertise in biological nomenclature is limited. Such knowledge can, however, be accurately modeled and delivered through a networked semantic resolution service that can provide end-users of biological nomenclatures or other dynamic terminologies with the appropriate information, in proper context, on demand. Such a service can also be used by database owners, publishers, or other information providers to semantically enable their offerings, making them more readily discoverable, even when the definition of a name or term has changed.

In our initial studies, we built a prototype to demonstrate that names, concepts, and the objects to which names apply must be treated independently in order to accurately interpret the correct meaning over time. As proof of principle, a preliminary data model and XML schema were developed and a simple semantic resolver was deployed that operated on an underlying dataset of bacterial and archaeal names. In Phase I, that data model was substantially refined to address limitations of the initial prototype and the underlying nomenclatural data were extensively cleaned, extended, and ported to a stable database environment to support development of commercial applications and services. In collaboration with the International Committee on the Systematics of Prokaryotes and the editorial office of the *Society for General Microbiology*, we built a prototype application of N4L to demonstrate feasibility of high-throughput semantic enablement of scientific literature at the pre-publication stage. An unanticipated outcome of that work was a realization that these same tools could also be used to also trap a wide variety of nomenclatural errors that would otherwise appear in print and in digital form.

Under a Phase II award, we are now extending the scope of data curation and building a framework for extending and distributing N4L information services to users of different classes. The NamesforLife database has undergone further refinement to improve the accuracy of the underlying data. Among the problems that remained at the conclusion of Phase I were a number of long-standing errors that have accumulated in the literature and public databases that were attributable to practices that did not conform with the rules of nomenclature or arose through the continual repropagation of errors in both data sources. Wherever possible, we have reviewed the original published sources to confirm the original observations. Objects in the database now retain the source from which the information was taken from. The NamesforLife data model was also modified to correctly model orthographic corrections, automatically created subspecies names (as per Rule 40d of the International Code of Bacterial Nomenclature) and emendations of taxon descriptions.

Web-based tools have been developed to facilitate data entry and retrieval by NamesforLife curators. In addition to entity editing, there are several ways to query for information

including by name, accession number, and a citation matcher to find references currently in the database. The curatorial tools connect to the database through a set of data validators to help curtail certain data entry errors such as ensuring exemplars are only connected to species/subspecies entities. As of December 2008, the database held records on 11,407 named bacterial and archaeal taxa (8732 species, 491 subspecies, 2184 higher taxa), along with 7747 records identifying verified specimen holdings in biological resource centers and links to 7365 references in which the related taxonomic and nomenclatural acts were effectively and validly published.

The target audience of NamesforLife services is the broad scientific community and others who may need-to-know the precise meaning of biological names or other terms, in correct temporal context as they are encountered in other digital content (scientific or technical literature, regulatory literature, databases, etc). The dynamic, yet asynchronous nature of biological nomenclature and similar terminology poses a significant burden on information providers as they must either invest in constantly maintaining their offerings to keep current or shift that burden to their end-users. If the former, the costs can be significant and, absent a means of synchronizing updates across an entire domain of knowledge, end users are still confronted with apparent discrepancies across data sources and content providers. If the burden is shifted to end-users, they must then locate alternative information sources, typically hosted through a web portal, that must be queried separately. This makes utilization of content cumbersome and can lead to considerable ambiguity.

The NamesforLife approach is to semantically enable content in a manner that is transparent to end-users at two points in the value chain: at the source (the data provider or publisher) and at the client side (the end-user). In either case, the end-user experience is the same. At each occurrence of a validly published bacterial or archaeal name, they can have access to precise authoritative information by simply clicking on the name. Tools to enable publishers content at the prepublishing stage by embedding persistent N4L identifiers into their content ensure that their readers will always have access to the correct meaning of the name (as well as additional information), even if the name has changed since publication. A web-based client has also been developed that supports semantic enablement of other digital content, one-the-fly, providing similar seamless access to NamesforLife content at each point where a name occurs. Collaborative prototyping of these services with early adopters will be discussed.

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

Standards in Genomic Sciences: an Open-Access, Standards-Supportive Publication that Rapidly Disseminates Concise Genome and Metagenome Reports in Compliance with MIGS/MIMS Standards

Scott H. Harrison^{1*} (harris41@msu.edu), Sam V. Angiuoli,² Patrick S.G. Chain,^{1,3,4,5} Dawn Field,⁶ Frank-Oliver Glöckner,⁷ Lynette Hirschman,⁸ Eugene Kolker,^{9,10} Nikos Kyrpides,⁴ Susanna-Assunta Sansone,¹¹ Lynn M. Schriml,² Peter Sterk,^{6,11} David W. Ussery,¹² Owen White,² and **George M. Garrity**^{1,5}

¹Microbiology & Molecular Genetics, Michigan State University, East Lansing, Mich.; ²Institute for Genome Sciences and Dept. of Epidemiology and Preventive Medicine, University of Maryland School of Medicine, Baltimore, Md.; ³Biosciences Directorate, Lawrence Livermore National Laboratory, Livermore, Calif.; ⁴DOE Joint Genome Institute, Walnut Creek, Calif.; ⁵Center for Microbial Ecology, Michigan State University, East Lansing, Mich.; ⁶NERC Centre for Ecology and Hydrology, Oxford, United Kingdom; ⁷Microbial Genomics Group, Max Planck Institute for Marine Microbiology, Bremen, Germany; ⁸Information Technology Center, The MITRE Corporation, Bedford, Mass.; ⁹Seattle Children's Hospital Research Institute, Seattle, Wash.; ¹⁰Division of Biomedical and Health Informatics, Dept. of Medical Education and Biomedical Information, University of Washington, Seattle, Wash.; ¹¹EMBL—European Bioinformatics Institute, Cambridge, United Kingdom; and ¹²Center for Biological Sequence Analysis, The Technical University of Denmark, Lyngby, Denmark

Project Goals: The Standards in Genomic Sciences eJournal (SIGS; <http://standardsingenomics.org>) is a newly established standards-supportive publication being developed to report on the exponentially increasing volume of genomic and metagenomic data. As the first standards-based journal in the life sciences, SIGS will provide immediate open access to its content from the outset, on the principle that making research freely available to the public supports a greater global exchange of knowledge to better accelerate scientific discovery.

The Standards in Genomic Sciences eJournal (SIGS; <http://standardsingenomics.org>) is a newly established standards-supportive publication being developed to report on the exponentially increasing volume of genomic and metagenomic data¹. SIGS will also include coverage of detailed standard operating procedures, meeting reports, reviews and commentaries, data policies, white papers and other gray literature that are relevant to genomic sciences, but absent from the scholarly literature. To achieve an objective of standards compliance, SIGS has been designed to use the specification for the minimum information about a genome

sequence (MIGS)². MIGS is developed and maintained by the Genomic Standards Consortium (GSC), a group formed to promote the development of standardized annotations of genomic investigations. As a specification and checklist, MIGS has helped inspire formal models for gathering information routinely included in primary publications such as environmental context, biotic relationship, relationship to oxygen, source material identifiers, nucleic acid sequence and sample metadata as well as the overall sequencing assay and genome annotation protocol. MIGS has been developed to address the genomic information of bacteria, archaea, eukaryotes, plasmids, viruses, and organelles. MIGS has been extended into a specification of minimum information about a metagenome sequence (MIMS). SIGS is the first standards-based journal in the life sciences and will provide immediate open access to its content from the outset, on the principle that making research freely available to the public supports a greater global exchange of knowledge to better accelerate scientific discovery.

SIGS presents new opportunities to tightly integrate biological data and peer-reviewed content. Where the GSC and other standards initiatives have sought to develop interoperable approaches to ensure consistency in semantic and syntactic annotation of genomes and metagenomes^{3,4}, SIGS will apply and extend these standards to published. While data sets may continue to grow, new goals for research emerge and database structures and applications change, SIGS will provide static, archival snapshots of genomic data and metadata as points of record that are enriched with interpretative commentary and authenticated by peer review and formal validation against specific versions of the defined MIGS/MIMS checklists. The usage of MIGS/MIMS checklists will help guide SIGS as a vehicle for rapidly publishing concise, highly structured short reports of sequenced genomes so that readers can readily make comparisons across taxa, and link out from these comparisons to knowledge existing elsewhere in the literature. Automated scoring will complement the peer review process to address coverage of standards² as implemented in minimal and extended forms of XML-based tagging approaches such as can be done for MIGS/MIMS with the Genomic Contextual Data Markup Language (GCDML)³. Furthermore, authors may use the opportunity of publishing in SIGS to make further claims concerning their reported data set in terms of how it may be a new finding for a novel clade or niche. As an open access journal, SIGS will be distributed to readers at no cost, with publication costs initially being absorbed through grants from the Michigan State University Foundation and the U.S. Department of Energy (DE-FG02-08ER64707). The estimated throughput for the initial year of publication in 2009 will be 200 articles, with a two to three week peer-review cycle.

Currently, information for over 50 short genome report articles has been provided to SIGS, and archival draft mockups have been marked up in XML based on the with the NLM DTD (<http://dtd.nlm.nih.gov>) for both short genome reports and short metagenome reports. SIGS has been registered with an ISSN number and has attained CrossRef membership so that digital object identifiers

(DOIs) can be assigned to each published manuscript. The editorial workflow and website for SIGS has been deployed with the Open Journal Systems (OJS) software¹, and OJS review forms have been customized to address MIGS/MIMS compliance for some of the organism types that will be reported. Specifically, for the benefit of those contributing content, SIGS has been designed to present authors with example manuscripts based on uploaded annotations of MIGS/MIMS-compliant data during the initial stages of the editorial workflow. This allows for flexibility where authors can begin submission with either weakly or fully populated grids of MIGS/MIMS compliant metadata and update a previously generated document with additional metadata, comments, and insights concerning a reported genome or metagenome. Especially in the case of metagenomics, it is expected that each aspect of an associated biosystem may not fully correspond to the simple content model of the standards specification. Comments by the author can serve to address this for consideration by peer review in the journal. We will develop automated services that will be integrated throughout the editorial workflow to simplify the process of standards compliance and generation of open, searchable, peer-reviewed content.

References

1. Garrity, G. M., Field, D., Kyrpides, N., Hirschman, L., Sansone, S. A., Angiuoli, S., Cole, J. R., Glöckner, F. O., Kolker, E., Kowalchuk, G., Moran, M. A., Ussery, D., and White, O. 2008. Toward a standards-compliant genomic and metagenomic publication record. *OMICS* 12:157-60.
2. Field, D. et al. 2008. The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* 26:541-7.
3. Kottmann, R., Gray, T., Murphy, S., Kagan, L., Kravitz, S., Lombardot, T., Field, D., Glöckner, F. O., and the Genomic Standards Consortium. 2008. A Standard MIGS/MIMS Compliant XML Schema: Toward the Development of the Genomic Contextual Data Markup Language (GCDML). *OMICS* 12:115-121.
4. Hirschman, L., Clark, C., Cohen, K. B., Mardis, S., Luciano, J., Kottmann, R., Cole, J., Markowitz, V., Kyrpides, N., Morrison, N., Schriml, L. M., Field, D. and the EnvO Project. 2008. Habitat-Lite: A GSC Case Study Based on Free Text Terms for Environmental Metadata. *OMICS* 12:129-136.

—
GTL

The Ribosomal Database Project: Tools and Sequences for rRNA Analysis

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

Michigan State University, East Lansing, Mich.

Through its website (<http://rdp.cme.msu.edu>), the Ribosomal Database Project (RDP) offers aligned and annotated rRNA sequence data and analysis services to the research community (Cole et al., 2008). These services help research-

ers with the discovery and characterization of microbes important to bioenergy production, biogeochemical cycles, and bioremediation.

Updated monthly, the RDP maintained 715,637 aligned and annotated quality-controlled rRNA sequences as of December 2008 (Release 10.6; Fig. 1). The *my*RDP features have grown to support a total of over 2,600 active researchers using their *my*RDP accounts to analyze over 2,000,000 pre-publication sequences in 22,721 sequence groups.

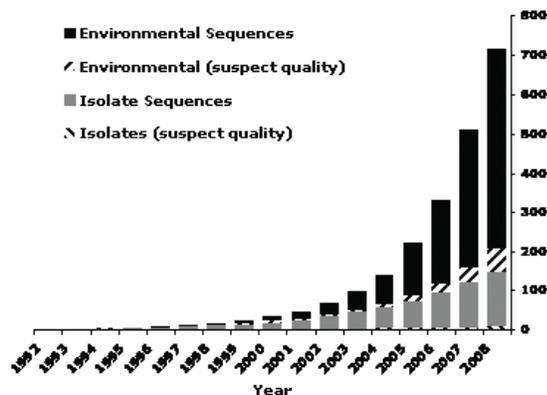


Figure 1. Increase in number of publicly available Bacterial and Archaeal small-subunit rRNA sequences. Suspect quality sequences were flagged as anomalous by Pintail in testing with two or more reference sequences from different publications.

New Bacterial and Archaeal Alignments: A major update to the RDP data sets was released in May 2008. Unlike the RDP 9 series of releases, the RDP 10 series provides an up-to-date aligned and annotated Archaeal data set along with the Bacterial data set. In addition, RDP 10 series alignments are created using the INFERNAL secondary structure based aligner (Nawrocki and Eddy, 2007). Both INFERNAL and RNACAD, our previous aligner, are based on stochastic context-free grammars and provide a high-quality secondary-structure aware alignment. The INFERNAL aligner provides several significant advantages over RNACAD. The INFERNAL aligner is about 25 times faster; it provides a much more intuitive handling of sequencing errors, and solves some known problems with incorrect alignment of short partial sequences. We trained the INFERNAL aligner on a small hand-curated set of high-quality full-length rRNA sequences derived mainly from genome sequencing projects. A relatively small training set of under 1000 representative sequences kept the alignment size small enough for reliable hand-adjustment of homology information in the training set. All *my*RDP users' sequences have been upgraded to the RDP 10 alignments. By adding Archaea, the RDP is responding to the needs of our user community. In addition, this release incorporates recent reevaluations to the *Firmicutes* and *Cyanobacteria* proposed by Bergey's Trust, along with additional published informal taxonomies for the *Acidobacteria*, *Verrucomicrobia*, OP11, and other less-well-studied areas of microbial diversity.

New High-Throughput Pyrosequencing Analysis Pipeline: In May 2008 the RDP released a new pipeline that provides tools to support analysis of next-generation ultra high-throughput rRNA sequencing data. This pipeline offers a collection of tools that automate the data processing and simplify the computationally intensive analysis of large sequencing libraries.

In the initial processing steps, raw sequence reads from multiple samples are sorted using sample-specific tag sequences. The mapping between the tag sequence and sample name is designated in a tag file. Four quality filters can be applied in the initial processing step. For taxonomy-independent alignment, the trimmed reads are aligned using the fast INFERNAL aligner trained on a small, hand-curated set of high-quality full-length rRNA sequences derived mainly from genome sequencing projects. Reads are then clustered into Operational Taxonomic Units at multiple pairwise distances using custom code implementing the complete-linkage clustering algorithm. Specialized tools provide common ecological metrics including: Chao1, Shannon Index and rarefaction. In addition, the processed data can be downloaded in formats suitable for common ecological and statistical packages including SPADE, EstimateS and R. Other options are available to cluster data from multiple samples, to combine alignments, to extract specific sequences from the dataset, to select representative sequences from clustered sequences and to produce comparative metrics among samples.

Several existing RDP tools are used for taxonomy-based analysis. The RDP Classifier provides fast and reliable classification of short sequence reads. The RDP Library Compare program can be used to detect differentially represented taxa between samples, and the RDP Sequence Match tool can be used to find the closest sequences in the RDP database for each sequence in a sample.

New "Taxomatic" Visualization Tool implementing SOSCC: This tool displays a color heat map representation of distances between large sets of sequences. It helps spot errors in the underlying taxonomy. Users can select a set of RDP and/or *my*RDP sequences and display the sequences in taxonomic order in Taxomatic. Thousands of sequences can be visualized at one time, or users can pan and zoom to examine individual sequences. Taxa that are phylogenetically incoherent or are misplaced stand out visually in the representation, as do individual sequences. Any taxonomic group or individual sequence can be highlighted, and tooltips are displayed upon mouseover to display information about individual sequences. The SOSCC algorithm provides a supervised reclassification of sequences and reordering of the distance matrix. The Taxomatic and SOSCC leverages work originally funded in a separate DOE grant.

New Web Services: The RDP offers new SOAP web services interfaces for the RDP SeqMatch and Classifier tools. These web services have been tested with Java and Ruby. Researchers can incorporate these web services to their own analysis pipelines to make use of these popular RDP tools.

New RDP Class Assignment Generator: This new educational tool for professors or instructors provides a lesson plan

along with individualized material. This generates unique sequences that can be easily distributed to a classroom, providing easy-to-follow instructions for students, and providing an answer key to evaluate the performance of students. The instructor is walked through a simple form that asks about the number of students, the level of difficulty (the number of sequences to be assigned for each student), and other information about the class. The tool then produces a set of sequences for each student modified from existing sequences. The modifications follow evolutionary principles and conserve the rRNA secondary structure. Since each student receives a customized set of sequences, there is less chance of students sharing results. The students are asked to analyze the sequences, and the instructor is provided with a key containing the correct classification for each student's sequences.

RDP User Surveys: Over the last year the RDP implemented a new user survey system to help obtain user input on directions for the RDP. Each survey asks a single question and is displayed for approximately two weeks as an overlay when users access the RDP website. Users may either answer the question or decline to answer before continuing to the RDP website. A browser cookie is used to keep users from seeing the same survey question twice. The results of the RDP surveys are available on the RDP website (<http://rdp.cme.msu.edu/misc/surveys.jsp>).

References

1. Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje (2008). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* doi:10.1093/nar/gkn879.
2. Nawrocki, E.P. and Eddy, S.R. (2007). Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput. Biol.*, 3:e56.

This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy under Contract No. DE-FG02-99ER62848.