DOE/SC-0089

DOE GENOMICS:GTL

ACCELERATING DISCOVERY FOR ENERGY AND ENVIRONMENT

OFFICE OF SCIENCE
U.S. DEPARTMENT OF ENERGY

Contractor-Grantee Workshop III

Washington, D.C.

February 6–9, 2005

Prepared for the
U.S. Department of Energy
Office of Science
Office of Biological and Environmental Research
Office of Advanced Scientific Computing Research
Germantown, MD 20874-1290

Prepared by
Genome Management Information System
Oak Ridge National Laboratory
Oak Ridge, TN 37830
Managed by UT-Battelle, LLC
For the U.S. Department of Energy
Under contract DE-AC05-00OR22725

Contents

Welcome to Genomics:GTL Workshop III

Genomics:GTL Program Projects

Harvard Medical School

Lawrence Berkeley National Laboratory

* Presenting author

Oak Ridge National Laboratory and Pacific Northwest National Laboratory

Poster Page

Sandia National Laboratories

* Presenting author

University of Massachusetts, Amherst

Shewanella Federation

* Presenting author

J. Craig Venter Institute

* Presenting author

Communication

Bioinformatics, Modeling, and Computation

* Presenting author

Environmental Genomics

* Presenting author

Microbial Genomics

* Presenting author

Technology Development and Use

Imaging, Molecular, and Cellular Analysis

* Presenting author

Protein Production and Molecular Tags

* Presenting author

Proteomics and Metabolomics

* Presenting author

Ethical, Legal, and Societal Issues

* Presenting author

Welcome to Genomics:GTL Workshop III

Welcome to the third Genomics:GTL Contractor-Grantee workshop. GTL continues to grow—scientifically, in DOE relevance, and as a program that needs all your diverse scientific, technical, and intellectual efforts to make it a success. GTL is attracting broad and enthusiastic interest and support from scientists at universities, national laboratories, and industry; colleagues at other federal agencies; Department of Energy leadership; and Congress.

GTL's challenge to the scientific community is to further develop and use a broad array of innovative technologies and computational tools to systematically leverage the knowledge and capabilities brought to us by DNA sequencing projects. The goal is to seek a broad and predictive understanding of the functioning and control of complex systems in individual microbes and microbial communities. GTL's prominent position at the interface of the physical, computational, and biological sciences is both a strength and a challenge. Microbes remain GTL's principal biological focus. In the complex "simplicity" of microbes, we find capabilities needed by DOE and the nation for clean and secure energy, cleanup of environmental contamination, and sequestration of atmospheric carbon dioxide that contributes to global warming. An ongoing challenge for the entire GTL community is to demonstrate that the fundamental science conducted in each of your research projects brings us a step closer to biology-based solutions for these important national energy and environmental needs.

This year brings two important milestones for GTL. First is the development of a roadmap that will help guide and justify the GTL program to a broad audience of scientists, policymakers, and the public. In the coming weeks we will be calling on many of you to provide critical review of this important document. Second is an important step forward in developing GTL user facilities: we are beginning the process of engineering and designing the Facility for Production and Characterization of Proteins and Molecular Tags.

GTL workshops are high-energy events that provide an opportunity for all of us to discuss, listen, and learn about exciting new advances in science; identify research needs and opportunities; form research partnerships; and share the excitement of this program with the broader scientific community. We look forward to a stimulating and productive meeting and offer our sincere thanks to all the organizers and to you, the scientists, whose vision and efforts will help us all to realize the promise of this exciting venture.

AA Patrinos

Ari Patrinos
Associate Director of Science for
Biological and Environmental Research
Office of Science
U.S. Department of Energy
ari.patrinos@science.doe.gov

Carl Edward Oliver

Ed Oliver
Associate Director of Science for
Advanced Scientific Computing Research
Office of Science
U.S. Department of Energy
ed.oliver@science.doe.gov

Genomics:GTL Program Projects

Harvard Medical School

1

Metabolic Network Modeling of *Prochlorococcus marinus*

George M. Church* (g1m1c1@arep.med.harvard.edu), Xiaoxia Lin, Daniel Segrè, Aaron Brandes, and Jeremy Zucker

Harvard Medical School, Boston, MA

The marine cyanobacterium *Prochlorococcus marinus* dominates the phytoplankton in the tropical and subtropical oceans and contributes to a significant fraction of the global photosynthesis. Several strains of *Prochlorococcus* have been sequenced, which provides us a promising starting point for investigating the relationship between genotype and phenotype at a genome scale and with a comparative approach. To achieve the ultimate goal of understanding the metabolism at a systems level, we are developing and utilizing new metabolic network models in several directions.

Comparison and connection of day-night metabolisms

Day-night cycles are known to play a central role in the metabolism of *Prochlorococcus*. We are exploring two approaches to model the difference and connection between day and night. One is to take the full metabolic network and formulate two separate models assuming different nutrient conditions and optimality criteria. Then the flux predictions can be compared to mRNA and protein expression data. In the other approach, we make use of the protein expression data, which helps to reduce the feasible flux space and leads to finer flux predictions.

Construction of metabolic networks

One major challenge in constructing complete and accurate *in silico* metabolic networks for quantitative analysis such as flux balance analysis (FBA) is to identify reactions that are "missed" in the annotation. We have been mainly using Pathway Tools software suite developed by SRI to identify metabolic reactions and are developing new algorithms to construct the "functional" metabolic network from a network perspective. Biochemical reactions with identified enzymes are included and then an "optimal" set of reactions are added such that the network produces the specified growth phenotype given corresponding nutrient conditions. Identification of the missing links will also help to refine the genome annotation. Another problem is that there exist "orphaned enzymes"— experimentally elucidated biochemical reactions whose enzyme has never been sequenced. To address this problem, we are utilizing a pathway hole-filling algorithm developed by SRI and developing bioinformatics techniques to identify candidate genes for these orphaned enzymes.

Analysis of metabolic networks with mass balance and energy balance

Conventional flux balance analysis (FBA) only considers mass balance. We are incorporating constraints representing the second law of thermodynamics, which eliminates thermodynamically infea-

* Presenting author

sible fluxes. A subset of the additional constraints exhibits non-convexity, giving rise to substantial difficulty in the solution of the resulting optimization problem. We are developing new methods to overcome this challenge to make full use of combined FBA and EBA (energy balance analysis).

Construction and comparative study of whole-cell metabolic networks of MED4 and other strains

By combining a bioinformatics pipeline for generating metabolic network models from genome annotations and manual inspection/modification, we have constructed the *in silico* metabolic network of central carbon metabolism and amino acid biosynthesis for *Prochlorococcus* MED4, a high-light-adapted strain. We are extending it towards the genome-wide network. In addition, we will construct metabolic network models for the other sequenced strains, including the low-light-adapted MIT9313. Comparison of the structures of their metabolic networks and the calculated flux distributions under varying conditions will enable us to understand at a systems level how these different strains adapt their metabolisms to the different environments.

Project Web site: http://arep.med.harvard.edu/DOEGTL/

2

Quantitative Proteomics of *Prochlorococcus marinus*

Kyriacos C. Leptos[1]* (leptos@fas.harvard.edu), Jacob D. Jaffe[1], Eric Zinser[2], Debbie Lindell[2], Sallie W. Chisholm[2], and George M. Church[1]

[1]Harvard Medical School, Boston, MA and [2]Massachusetts Institute of Technology, Cambridge, MA

With the capability of performing whole-cell proteome analysis, a need to extent the above capability to whole-cell protein quantitation has proven to be a necessity. For this purpose we developed MapQuant, a platform-independent open-source software, which given large amounts of mass-spectrometry data, outputs quantitation for any organic species in the sample. We have previously applied MapQuant in the study of standardization samples at different concentrations on both LCQ and LTQ-FT spectrometers and also in the content of protein mixture of medium complexity and have showed linearity of signal with respect to the quantity of protein introduced.

The *Prochlorococcus* species is an abundant marine cyanobacterium that contributes significantly to the primary production of the ocean and whose life cycle is synchronized to the solar day (the “diel cycle”). In this study we leverage previously obtained protein identification data and the capabilities of MapQuant to quantify the proteins in a time-series dataset which includes 25 time points distributed along a 48-hour period (two diel cycles) of the strain MED4 of *Prochlorococcus marinus*. Protein samples from the growing culture were collected in duplicate and digested into peptides using trypsin, each time-point sample subjected to liquid chromatography coupled to hybrid linear ion trap-FTICR mass spectrometry, giving rise to a total 150 LC/MS experiments. The data acquisition took place on a Finnigan LTQ-FT mass spectrometer and it involved the acquisition of maximum two MS/MS spectra per MS spectrum. MS/MS spectra were interpreted using the program SEQUEST. The cross-correlation scores assigned to peptides that scored were filtered using thresholds to take into account false-positive results and the peptides were compiled into a summary list. This list of highly scored peptides was used as landmarks for evaluating MapQuant performance. MapQuant algorithms include morphological operations, noise filtering, watershed segmentation, peak finding and fitting, peak clustering and isotopic-cluster deconvolution and fitting using binomially distributed clusters of gaussioid peaks.

* Presenting author

MapQuant outputs a list of potential organic species, by reporting four physical attributes for each isotopic cluster that it deconvolves. Those attributes are the m/z and the retention time (RT) of the monoisotopic peak, its charge and its carbon content. We have employed an m/z, RT and charge matching approach to assigning MapQuant Isotopic Clusters (MQIC) to the landmark peptides identified by SEQUEST in the same run with 91% success. However, MQICs that were assigned to a peptide using SEQUEST constitute 3% of the total MQIC found in a 2-D map. We are in the process of developing a matching algorithm that will be able to assign identities to unassigned MQICs. This approach will utilize SEQUEST peptides identified in the same organism *Prochlorococcus marinus* MED4 in five LC/LC/MS/MS experiments performed in the past, which correspond to five different environmental conditions. The matching algorithm should enable mapping of many of the remaining (97%) of the unidentified MQICs.

Our end goal is to be able to perform quantitation for most peptides found in the 25 time-points of the two diel cycles and hope to understand how carbon fixation, light-response and cell division are coordinated throughout the daily cycle.

Project Web site: http://arep.med.harvard.edu/DOEGTL/

3

Genome Sequencing from Single Cells with Ploning

Kun Zhang[1]* (kzhang@genetics.med.harvard.edu), Adam C. Martiny[2], Nikkos B. Reppas[1], Sallie W. Chisholm[2], and George M. Church[1]

[1]Harvard Medical School, Boston, MA and [2]Massachusetts Institute of Technology, Cambridge, MA

Currently genome sequencing is performed on cell populations because of the difficulty in preparing sequencing template from single cells. This makes the genome sequences of many difficult-to-culture organisms inaccessible or poorly assembled. We have developed a method that enables genome sequencing from a single cell by performing polymerase cloning (ploning). In this method, we prepare sequencing templates from single cells with real-time multiple displacement amplification (rtMDA), which allows us to tackle the big technical challenge in single-cell whole genome analysis: to detect and suppress spurious amplification while targeting a single molecule of a microbial chromosome.

Experiments on *Escherichia coli* show that, (1) an amplification magnitude of 10^9 was achieved by rtMDA, (2) strain-specific genetic signatures were preserved, (3) neither spurious amplification product nor chimeric sequence was detected, (4) an estimated 97% of the target genome could be recovered from a polymerase clone (plone) at the 10X sequencing depth. The remaining regions are not missing, but present at lower copy numbers, and easily recovered by PCR. Since the low-coverage regions seem random, genome coverage can be improved by pooling the sequencing reads from two or more plones of the same type of cells during the assembly stage. Furthermore, we successfully performed ploning on both fresh and frozen *Prochlorococcus* cells, and obtained nearly complete coverage on both strains (MED4 and MIT9312) we tested. Plones of single cells from an ocean sample (from the Hawaii Ocean Time-series) are being screened for *Prochlorococcus* cells for genome sequencing. Initial results indicate successful amplification of single *Prochlorococcus* cells from this sample. After further screening of genome coverage, whole genome shot-gun sequencing will be performed on a few selected plones.

* Presenting author

Lawrence Berkeley National Laboratory

4

VIMSS Computational Microbiology Core Research on Comparative and Functional Genomics

Adam Arkin*[1,2,3] (aparkin@lbl.gov), Eric Alm[1], Inna Dubchak[1], Mikhail Gelfand, Katherine Huang[1], Vijaya Natarajan[1], Morgan Price[1], and Yue Wang[2]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]University of California, Berkeley, CA; and [3]Howard Hughes Medical Institute, Chevy Chase, MD

Background. The VIMSS Computational Core group is tasked with data management, statistical analysis, and comparative and evolutionary genomics for the larger VIMSS effort. In the early years of this project, we focused on genome sequence analysis including development of an operon prediction algorithm which has been validated across a number of phylogenetically diverse species. Recently, the Computational Core group has expanded its efforts, integrating large amounts of functional genomic data from several species into its comparative genomic framework.

Operon Prediction. To understand how bacteria work from genome sequences, before considering experimental data, we developed methods for identifying groups of functionally related genes. Many bacterial genes are organized in linear groups called operons. The problem of identifying operons had been well studied in model organisms such as *E. coli*, but we wished to predict operons in less studied bacteria such as *D. vulgaris*, where data to train the prediction method is not available. We used comparisons across dozens of genomes to identify likely conserved operons, and used these conserved operons instead of training data. The predicted operons from this approach show good agreement with known operons in model organisms and with gene expression data from diverse bacteria.

Statistical Modeling of Functional Genomics Experiments. These operon predictions give hints to the function and regulation of many genes, but they can also aid the analysis of gene expression data. Genes in the same operon generally have similar expression patterns, so the degree to which genes in the same operon have correlated measurements gives an indication of the reliability of the data. Although most analyses of gene expression data have assumed that there are no systematic biases, we found that many data sets have systematic biases – biases that can not be corrected simply by increasing the number of experimental replicates. Using *a priori* knowledge of operon structure from our predictions, we can measure and account for these systematic biases, and more accurately assign confidence levels to experimental measurements. Furthermore, if several genes in an operon have consistent measurements, we have developed novel statistical models that assign much higher confidence to those measurements.

Evolution of Microbial Genomes. Our analysis of operons also led us to discoveries about how bacteria evolve. First, a popular theory has been that operons are assembled by horizontal gene transfer, and that operons exist, in part, to facilitate such transfers. We showed that such transfers are not involved in operon formation, and instead argue that operons evolve because they improve gene regulation. Second, we discovered that operons are preferentially found on the leading strand of DNA replication. (In most bacteria, a majority of genes are on the leading strand.) This observation

* Presenting author

is not explained the leading theories for strand bias. Instead, we note that genes, and especially long operons, are turned off during DNA replication, and these disruptions are shorter for operons on the leading strand. We believe that this mechanism can explain the known patterns of strand bias.

Metabolic Reconstruction of Delta-Proteobacteria. Species in the delta subgroup of the proteobacteria represent an important constituent of natural environmental diversity with key properties such as the ability to reduce heavy metals that make them of particular relevance to DOE core missions. Recently, a number of delta-proteobacterial genomes were sequenced, yet little is known about the physiology and regulation of key pathways. We have completed a comprehensive survey of regulatory signals and metabolic reconstruction of metal-reducing delta-proteobacterial species using comparative genomic analysis. In our survey, we characterized the evolution of 15 distinct regulons across six species. Interestingly, these species shared as many regulatory pathways in common with *B. subtilis*, a gram-positive bacterium, as they did with *E. coli*, itself a member of the proteobacteria. In addition to previously characterized regulons, we discovered a new CRP-like transcription factor that controls the sulfate-reduction machinery in *Desulfovibrio spp.*, and is generally present across anaerobic species, which we have named HcpR.

Data Analysis. The Computational Core group also played a role in the interpretation of experimental data generated by the Functional Genomics Core group. In a recent experiment in which *D. vulgaris* cells were subject to nitrite stress, the Computation Core group developed a detailed biological model that explains the observed transcriptional responses at a molecular level. In particular, enzymes involved in nitrite reduction to ammonia and incorporation of ammonia into glutamate were up-regulated, while the sulfate reduction machinery was down-regulated. In addition, iron uptake and oxidative stress genes were found to be up-regulated. Individual transcription factors along with their cognate DNA motifs were identified for each of these responses, and a model was proposed in which nitrite or other nitrogen intermediates play a role in oxidizing Fe(II), which in turn de-represses transcription from both the iron uptake and oxidative stress regulons.

Data Management. To support the larger VIMSS effort, the computational core group has deployed several new databases: the Biofiles database for rapid upload of arbitrary data types; the Experimental Data Staging and Experiment/Data Reporting Systems (EDSS/EDR) to automate the processing of key data types such as gene expression experiments; and the MicrobesOnline database which features a suite of analysis and visualization tools.

The EDSS database contains information and data from biomass production experiments (time points, stressor, direct cells counts, micrographs) and growth curve experiments. Several Web interfaces have been developed to access the EDSS database, including, details about the biomass production experiments (lab procedures, sample allocations, shipping conditions), tables of QA data (direct counts), and plots of growth curve data. In addition, time points and information about stressors stored in EDSS are accessed when the results of other experiments (e.g., microarray experiments) are analyzed and results compared. The EDR database and Web interface were developed to provide a reporting system to track data generation from the starting point of biomass production through the entire suite of laboratory analyses performed on the biomass. The reporting system allows PIs to document each step in the experimental pipeline (e.g., sample preparation, QA measurements, etc.). A major component of the EDR system is a Web interface for writing and submitting reports about data being uploaded to the VIMSS file server. The interface requires users to describe the laboratory analysis that generated the data (type of analysis, dates data were generated, biomass source, etc.), content of the uploaded data, the file format and the format of the data within the file(s), and any reference information needed to fully understand the data file(s).

The MicrobesOnline Database. The MicrobesOnline database currently hosts 180 genomes and features a full suite of software tools for browsing and comparing microbial genomes. Highlights include operon and regulon predictions, a multi-species genome browser, a multi-species Gene Ontology browser, a comparative KEGG metabolic pathway viewer and the VIMSS Bioinformatics Workbench for more in-depth sequence analysis. In addition, we provide an interface for genome annotation, which like all of the tools reported here, is freely available to the scientific community. To keep up with the ever-increasing rate at which microbial genomes are being sequenced, we have established an automated genome import pipeline. Since August 2004 this automated pipeline has allowed us to increase the number of hosted genomes from 100 to 180.

A number of outside groups are currently using the MicrobesOnline database for genome annotation projects. To facilitate the use of this community resource we are developing a sophisticated access control system, so individual research groups can use the power of the VIMSS annotation tools, while keeping data from their own particular genome project private until their analyses are ready to be made public.

Addition of Functional Genomics to MicrobesOnline. In addition to browsing comparative genomics, the MicrobesOnline database and website now allows users to browse and compare functional genomics data. In particular we have started with gene expression microarray data as a test case for high-throughput functional genomics measurements. Currently gene expression data from 262 experiments across four different species are hosted in the database. Software tools available from the MicrobesOnline functional genomics web portal allow users to overlay expression data on predicted operon structure or metabolic pathways. In addition, an operon-based estimate of microarray accuracy has proven useful in determining the quality of experimental measurements.

5

The Virtual Institute of Microbial Stress and Survival (VIMSS): Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Microbes

Carl Abulencia[4], Eric Alm[1], Gary Andersen[1], Adam Arkin[1]* (APArkin@lbl.gov), Kelly Bender[5], Sharon Borglin[1], Eoin Brodie[1], Swapnil Chhabra[3], Steve van Dien[6], Inna Dubchak[1], Matthew Fields[7], Sara Gaucher[3], Jil Geller[1], Masood Hadi[3], Terry Hazen[1], Qiang He[2], Zhili He[2], Hoi-Ying Holman[1], Katherine Huang[1], Rick Huang[1], Janet Jacobsen[1], Dominique Joyner[1], Jay Keasling[1], Keith Keller[1], Martin Keller[4], Aindrila Mukhopadhyay[1], Morgan Price[1], Joseph A. Ringbauer, Jr.[5], Anup Singh[3], David Stahl[6], Sergey Stolyar[6], Jun Sun[4], Dorothea Thompson[2], Christopher Walker[6], Judy Wall[5], Jing Wei[4], Denise Wolf[1], Denise Wyborski[4], Huei-che Yen[5], Grant Zane[5], Jizhong Zhou[2], and Beto Zuniga[6]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]Oak Ridge National Laboratory, Oak Ridge, TN; [3]Sandia National Laboratories, Livermore, CA; [4]Diversa, Inc., San Diego, CA; [5]University of Missouri, Columbia, MO; [6]University of Washington, Seattle, WA; and [7]Miami University, Oxford, OH

Introduction

The mission of the Virtual Institute of Microbial Stress and Survival, is to understand the molecular basis for the survival and growth of microbes in the environment. Towards this end VIMSS has

* Presenting author

designed a series of key protocols, experimental pipelines and computational analysis to support and coordinate research in this area. Our flagship project aims to elucidate the pathways and community interactions which underlie the ability of *Desulfovibrio vulgaris* Hildenborough (DvH) to survive in diverse, possibly contaminated environments and reduce metals. Their ability to reduce toxic Uranium and Chromium, major contaminants of industrial and DOE waste sites, to a less soluble form has made them attractive from the perspective of bioremediation.

We are discovering the molecular basis for the physiology of these organisms first through characterization of the biogeochemical environment in which these microbes live and how different features of these environments affect their growth and reductive potential. We have created an integrated program through the creation of an experimental pipeline for the physiological and functional genomic characterization of microbes under diverse perturbations. This pipeline produced controlled biomass for a plethora of analyses as described below and is managed through workflow tools and a data management and analysis system. The effort is broken into three interacting core activities: The Applied Environmental Microbiology Core; the Functional Genomics Core; and the Computational Core.

Accomplishments of the Applied Environmental Microbiology Core (AEMC)

Characterization of the Environment. The AEMC has collected or completed basic analysis of the stressors present at a number of NABIR FRC site, and characterized the microbial community before and after stimulation using 16SRNA microarrays. Large insert cloning was used to characterize the enrichment of genomic functions in these environments. Diversity analysis of library clones revealed genes used in transport, small molecule binding, toxicity response and DNA synthesis, among others. We are now targeting primers for enrichment of signal transduction pathway components. In addition, nine *D. vulgaris*-like bacteria (DP1-9) were isolated from a metal impacted field site (Lake DePue, Illinois). All had identical 16S rRNA and *dsrAB* genes that were virtually identical to the orthologous genes of DvH. Complementary whole-genome microarray hybridization revealed that approximately 300 deleted genes were distributed in six regions of the chromosome, annotated as conserved/ hypothetical or phage related genes in DvH. We are now following up characterization of these phageless strains.

Biomass Production and Characterization: In the core pipeline experiments each microbe is first characterized physiologically using Omnilog phenotypic microarrays. A stressor condition is then applied to a large set of batch cultures and samples are collected periodically to obtain a time-series of cellular response. Each time-point is split so that the cells can be imaged, analyzed through synchrotron IR microscopy to measure the bulk physiological changes of the cells during their response, and determine the optimal time points to send to the functional genomics core (FGC) for transcript, protein and metabolite analysis. Response to oxygen stress, salt stress (shock and adaptation) and nitrate have been fully characterized in this way. In related work, we are developing laboratory systems that simulate environmental conditions than can not be achieved in pure culture, initially focusing on co-cultures of two different *Desulfovibrio* species (*DvH* and *Desulfovibrio sp*. PT2) syntrophically coupled to a hydrogenotrophic methanogen (*Methanococcus maripaludis*). Transcriptional dynamics of the co-culture has been measured by the FGC. In addition, a metabolic stoichiometric model has been constructed using flux balance analysis (FBA) to complement and direct experimental studies on the physiology of *DvH* growing either alone or in co-culture.

Accomplishments of the Functional Genomics Core

Genetics: To improve the genetic accessibility of *DvH*, we found the cells to be sensitive to the antibiotic Geneticin or G418, therefore, allowing kanamycin resistance to be used as a genetic marker.

Using the modified mini-Tn5 from Bill Metcalf, we have been able to generate a library of transposon mutants that appear to be randomly inserted throughout the genome. Several putative regulatory genes were among those mutated and we are screening for mutants of specific phenotypes. We have generated tagged hspC and rpoB genes in single copy controlled by their native promoters to use for development of assays for protein complexes. We have established a procedure for making gene deletions in non-essential genes that introduces a unique oligonucleotide that can be used for mutant identification. With this procedure, we have generated a putative *fur* deletion that is increased four fold over the wild type in its resistance to manganese. We are also generating a library of histidine kinase (HK) knockouts. *DvH* has 69 HKs that govern signal transduction. A suicide vector has been designed and created to enable gene deletion and concurrent "bar-coding" of the chromosome. Our preliminary results include 6 potential knock-out mutants.

Transcriptomics: We have, to date, characterized five stresses in *DvH* and five in *S. oneidensis* and results are integrated with the VIMSS MicrobesOnline Database. New regulons and their cis-regulatory sequences have been discovered along with new hypotheses of the pathways by which both organisms respond to these different stressors. A number of papers are in press, submitted or are in preparation around this topic.

Proteomics: We have developed three complementary proteomics methods to characterize protein expression in our microbes Differential In Gel Electrophoresis followed by MALDI-TOF and nanLC-ISI-QTOF, Isotope coded affinity tagging with tandem LC mass spec, and direct MS-MS. In addition, to characterize protein complexes we have developed both a high throughput cloning & expression of DvH proteins in *E. coli* and methods for expression of genetically-modified proteins at their native levels in the host organism. These proteins are then used as bait proteins to enable "pull-down" of associated proteins.

Metabolomics: We have set up and optimized both Capillary electrophoresis (CE) and Liquid chromatography (LC) coupled with Mass spectrometry (MS) methods for characterization of metabolites. Metabolite extraction protocols have been developed for *DvH*.

Accomplishments of the Computational Core

During the past year the computational core has focused on building the comparative and functional genomic analysis tools to aid in the prediction of regulatory networks in microbes, elucidate their evolutionary relationships and extract the most meaning from the functional genomics and phenotypic data described in the last two sections. We have developed an increasingly sophisticated experimental and data management system that centralizes and serves all VIMSS data and tracks the progress through experimental runs of the pipeline. One of the key technologies we have developed is a set of web-accessible comparative genomic tools (http://vimss.org) designed to facilitate multi-species comparison among prokaryotes. Highlights of the system accessible through the VIMSS website include operon and regulon predictions based on novel methods we have proven to work on a wide diversity of micro-organisms, a multi-species genome browser, a multi-species Gene Ontology browser, a comparative KEGG metabolic pathway viewer and the VIMSS Bioinformatics Workbench for in-depth sequence analysis. In addition, we provide an interface for genome annotation, which like all of the tools reported here, is freely available to the scientific community. This tool has been used successfully by a number of projects. In particular, an Joint Genome Institute Annotation Jamboree we ran to annotate *D. desulfuricans* G20 which will likely be reclassified as *D. alaskensis*. We have also been working on tools for modeling pathways and understanding how the molecular strategies we measure in the lab confer the ability to survive in the environment.

* Presenting author

6

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

Terry C. Hazen*[1] (tchazen@lbl.gov), Carl Abulencia[3], Gary Andersen[1], Sharon Borglin[1], Eoin Brodie[1], Steve van Dien[5], Matthew Fields[6], Jil Geller[1], Hoi-Ying Holman[1], Rick Huang[1], Janet Jacobsen[1], Dominique Joyner[1], Martin Keller[3], Aindrila Mukhopadhyay[1], David Stahl[5], Sergey Stolyar[5], Jun Sun[3], Dorothea Thompson[2], Judy Wall[4], Denise Wyborski[3], Huei-che Yen[4], Grant Zane[4], Jizhong Zhou[2], and Beto Zuniga[5]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]Oak Ridge National Laboratory, Oak Ridge, TN; [3]Diversa, Inc., San Diego, CA; [4]University of Missouri, Columbia, MO; [5]University of Washington, Seattle, WA; and [6]Miami University, Oxford, OH

Field Studies

Identification of Different Relationships Between Contaminated Groundwater Samples Based Upon Geochemical Data or Multiple Gene Sequences from Microbial Communities. Factor analysis was used to identify a subset of variables that may explain a majority of the observed variance between the contaminated groundwater sites, and principal components analyses were then used to compare the sites based upon geochemistry, phylogenetic markers (n=353), and functional markers (n=432). The clonal libraries of the multiple genes (SSU rRNA gene, *nir*K, *nir*S, *amo*A, *pmo*A, and *dsr*AB) were constructed from groundwater samples (n=6) that varied in degrees of contamination. When geochemical characteristics were analyzed, the data suggested that the samples could be differentiated based upon pH, nitrate, sulfate, nickel, aluminum, and uranium. Similar relationships between the sites were observed when 107 analytes were used, but more resolution was achieved between the more contaminated sites. In addition, a majority of the variance between the acidic samples could be accounted for by tetrachloroethene, ^{99}Tc, SO_4, Al, Th and 1,1,2-trichloro-1,2,2-trifluoroethane. The analysis based on a phylogenetic marker resulted in different groupings for background and the two circumneutral sites compared to the geochemical analysis, and analyses of the OTU distributions for the functional genes each predicted different relationships between the sites. A tripartite PCA explained 76% of the variance and grouped the background sample with the three, heavily contaminated sites. When all gene OTUs were used in the analyses, the sites were more similar than in any other comparison, 94% of the observed variance cold be explained, the background site was grouped with the contaminated sites, and possible key populations were identified by factor analysis. The data suggested that even though the background site was phylogenetically and geochemically distinct from the acidic sites, the extreme conditions of the acidic samples might be more analogous to the limited-nutrient conditions of the background site.

Biopanning/Clone libraries. Diversa extracted high molecular weight DNA from organisms present in contaminated soil sediment samples using a method that preserves the integrity of the DNA. Because the number of organisms in these samples was low, the genomic DNA was amplified using a phage polymerase amplification system. 16S rRNA analysis was then used to examine the microbial diversity of the samples. The amplified DNA was also used in the construction of large and small insert DNA libraries. These libraries were then screened for the presence of histidine kinase genes with homology to a subfamily of *Desulfovibrio vulgaris* histidine kinases. Genomic DNA has been extracted and amplified from nine different sites at the NABIR field research center. 16S rRNA analysis revealed the presence of distinct bacterial phyla, including proteobacteria, acidobacteria,

* Presenting author

and planctomycetes. Small and large insert libraries were constructed for all samples and examined for clonal diversity. Plaque hybridization of these libraries to histidine kinase homologous probes resulted in multiple positive clones. These clones will be compared and used to develop a better understanding of cellular responses to different environmental factors. These experiments have furthered the understanding of how the biological organisms in a contaminated system are organized, regulated and linked.

Enrichments. Nine *D. vulgaris*-like bacteria (DP1-9) were isolated from a metal impacted field site (Lake DePue, Illinois) as an additional reference set for comparative stress analyses. All had identical 16S rRNA and *dsrAB* genes that were virtually identical to the orthologous genes of *D. vulgaris* Hildenborough (DvH). However, pulse field gel electrophoretic analysis of I-CeuI digests identified a large deletion in the genomes of all isolates. Complementary whole-genome microarray hybridization revealed that approximately 300 deleted genes were distributed in six regions of the chromosome, annotated as conserved/ hypothetical or phage related genes in DvH. These deletions were also confirmed by PCR analysis, using primers complementary to regions flanking the deletions. Continuing collaboration with Judy Wall (U Missouri) has shown that one of the "phage-deficient" *D. vulgaris* strains (DP4) serves as host for latent viruses of *D. vulgaris* Hildenborough, identifying two phage morphotypes by EM. MPN enrichments from FRC area 2 sediments were developed using a PIPES buffered B2 medium supplemented with: 1) lactate, 2) lactate plus ethanol, 3) acetate, 4) propionate 5) pyruvate or 6) hydrogen plus carbon dioxide. All showed sulfate reducing activity within a range of 10^{-1} to 10^{-4} dilutions. Thirty isolates from the lactate medium were shown by 16S rRNA sequence to be affiliated with the "*Firmicutes*". A Gram-negative sulfate reducer (curved-rod morphology) maintained on an H_2/CO_2 plus acetate medium was also isolated.

Dual culture systems. The kinetics and stoichiometry of syntrophic growth were determined in batch culture by quantifying each population, substrate consumption (lactate), evolution of metabolic intermediates (H_2 and acetate), and end-product accumulation (CO_2 and methane). *D. vulgaris* monocultures were grown at generation times comparable to syntrophic batch cultures (24 and 36 hours) in sulfate-limited chemostats for comparative transcription analyses. Fermentative growth *D. vulgaris* on a lactate medium (sulfate minus) with continuous headspace purging was also developed for comparison. Transcription analyses of co-cultures identified a preliminary set of *D. vulgaris* genes either up or down regulated with syntrophic association, including periplasmic and cytoplasmic hydrogenases. These analyses are now being replicated at ORNL. A metabolic stoichiometric model was constructed using flux balance analysis (FBA) to complement and direct experimental studies on the physiology of *D. vulgaris* growing either alone or in co-culture. The network for each organism was based primarily on the annotated genome sequences, supplemented by available biochemical knowledge. The *Desulfovibrio* model consists of 86 reactions and 73 internal metabolites, while that of the methanogen contains 84 reactions and 72 metabolites.

Stress Experiments

High Throughput Biomass Production. Producing large quantities of high quality and defensibly reproducible cells that have been exposed to specific environmental stressors is critical to high throughput and concomitant analyses using transcriptomics, proteomics, metabolomics, and lipidomics. Culture of *D. vulgaris* is made even more difficult because it is an obligate anaerobe and sulfate reducer. For the past two years, our Genomics:GTL VIMSS project has developed defined media, stock culture handling, scale-up protocols, bioreactors, and cell harvesting protocols to maximize throughput for simultaneous sampling for lipidomics, transcriptomics, proteomics, and metabolomics. All cells for every experiment, for every analysis are within two subcultures of the original ATCC culture of *D.*

* Presenting author

vulgaris. In the past two years we have produced biomass for 38 integrated experiments (oxygen, NaCl, NO_3, NO_2, heat shock, cold shock, pH) each with as much as 30 liters of mid-log phase cells (3 x 10^8 cells/ml). In addition, more than 40 adhoc experiments for supportive studies have been done each with 1-6 liters of culture. All cultures, all media components, all protocols, all analyses, all instruments, and all shipping records are completely documented using QA/QC level 1 for every experiment and made available to all investigators on the VIMSS Biofiles database (http://vimss.lbl.gov/perl/biofiles). To determine the optimal growth conditions and determine the minimum inhibitory concentration (MIC) of different stressors we adapted plate reader technology using Biolog and Omnilog readers using anaerobic bags and sealed plates. Since each well of the 96-well plate produces an automated growth curve, over more than 200 h, this has enabled us to do more than 4,000 growth curves over the last two years. Since the Omnilog can monitor 50 plates at a time, this allows us to do more than 5,000 growth curves in a year. We have also developed chemostat techniques using a specially made extremophile fermentor (FairManTech) that has no internal metal parts. With this system we can get *D. vulgaris* to steady state from the freezer in less than 80 h in turbidostat mode, with a dilution rate of 0.25 l/h. Each reactor has a useable volume of 3 liters, with our current two reactors this enables production of 6 liters of steady state culture twice a week. We have also developed new harvesting techniques to minimize the stress caused by sample preparation for shipping. Since the volumes being centrifuged are large, the cells were not cooling fast enough to ensure high quality samples, so we devised a sampling apparatus that draws the cells from the culture vessel through capillary tubing in a MgCl ice bath that lowers the sample to 4°C in less than 20 sec. These procedures have maximized our reproducibility and throughput for the 8 labs involved.

Phenotypic Responses. Phenotypic Microarray™ analysis is a recently developed analytical tool to determine the phenotype of an organism. The plates, which are commercially available from Biolog™ (Hayward, CA), consist of 20 96-well plates. The first eight plates test a variety of metabolic agents, including electron donors, acceptors, and amino acids. Plates 9 and 10 cover a pH and osmotic stressors, while plates 11-20 contain a variety of inhibitors, including toxic agents and antibiotics. We have developed the ability use these plates under anaerobic conditions by inoculating plates in an anaerobic chamber, and heat-sealing them in polyethylene bags containing an anaerobic sachet. Using this technique, anaerobic conditions were maintained for up to a week. It was found that preconditioning of the cells in specialized media was required for the different types of plates in order to get a valid phenotype. The plates have been successfully used to characterize the phenotype of the *D. vulgaris* ATCC strain and are currently being applied to mutant strains to provide rapid screening of mutant phenotypic changes, for rapid pathway analyses and modeling. See (https://vimss.lbl.gov/~jsjacobsen/cgi-bin/Test/HazenLab/Omnilog/home.cgi) for sample data sets and analyses.

Synchrotron FTIR Spectromicroscopy for Real-Time Stress Analysis. LBNL's newly developed synchrotron radiation-based (SR) Fourier-transform infrared (FTIR) spectromicroscopy beamline allows the study of many biochemical and biophysical phenomena non-invasively as the processes are happening. It has been enabling for our high throughput determinations of optimal sampling times stress experiments. We can observe real-time changes of major biomolecule pools within cells as they are exposed to different stressors. This allows us to pick optimal sampling points during the stress response for transcriptome, protemome, meatabolome, and lipidome analyses. It also allows us to verify the purity and state of the culture for QA/QC.

7

VIMSS Functional Genomics Core: Analysis of Stress Response Pathways in Metal-Reducing Bacteria

Aindrila Mukhopadhyay[1], Steven Brown[4], Swapnil Chhabra[2], Brett Emo[3], Weimin Gao[4], Sara Gaucher[2], Masood Hadi[2], Qiang He[4], Zhili He[4], Ting Li[4], Yongqing Liu[4], Alyssa Redding[1], Joseph Ringbauer, Jr.[3], Dawn Stanek[4], Jun Sun[5], Lianhong Sun[1], Jing Wei[5], Liyou Wu[4], Huei-Che Yen[3], Wen Yu[5], Grant Zane[3], Matthew Fields[4], Martin Keller[5] (mkeller@diversa.com), Anup Singh[2] (aksingh@sandia.gov), Dorothea Thompson[4], Judy Wall[3] (wallj@missouri.edu), Jizhong Zhou[4] (zhouj@ornl.gov), and Jay Keasling[1]* (keasling@socrates.berkeley.edu)

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]Sandia National Laboratories, Livermore, CA; [3]University of Missouri, Columbia, MO; [4]Oak Ridge National Laboratory, Oak Ridge, TN; and [5]Diversa Inc., San Diego, CA

The Functional Genomics Core is part of the VIMSS project and there is a separate overview presentation of the VIMSS project. Environmental contamination by metals and radionuclides constitutes a serious problem in many ecosystems. Bioremediation schemes involving dissimilatory metal ion-reducing bacteria are attractive for their cost-effectiveness and limited physical detriment and disturbance on the environment. *Desulfovibrio vulgaris, Shewanella oneidensis*, and *Geobacter metallireducens* represent three different groups of organisms capable of metal and radionuclide reduction whose complete genome sequences were determined under the support of DOE-funded projects. Utilizing the available genome sequence information, we have focused our efforts on the experimental analysis of various stress response pathways in *D. vulgaris* Hildenborough using a repertoire of functional genomic tools and mutational analysis.

Originally isolated in 1946, from the clay soils in Hildenborough, Kent (UK), *Desulfovibrio vulgaris* Hildenborough belongs to a class of sulfate reducing bacteria (SRB) that are found ubiquitously in nature. As with most soil bacteria that do not live permanently in hyperosmotic environments, a NaCl salt stress in a *D .vulgaris* can be expected to result in at least two primary responses, osmotic and that towards Na+ ions. The genomic sequence of *D. vulgaris* indicates that a variety of mechanisms may be employed to counter these two stresses. In order to understand these mechanisms at the physiological level an integrated functional genomics analysis was conducted. Data from microarray analysis of the transcriptome, quantitative and qualitative proteomic analysis, PLFA profiling and IR studies reveal many interesting responses to stress. For example, with respect to hyperosmotic stress, the three gene operon that regulates the uptake of the osmoprotectant, glycine betaine is highly up-regulated. With respect to Na+ stress, Na+/H+ antiporters such as the dehydrogenase *mnhA* show upregulation in mRNA levels. As might be expected with cellular physiology, a myriad of other relevant responses were observed such as upregulation in ATP synthesis, down-regulation in flagellar systems. IR studies also indicate changes in cell wall composition. Moreover, several genes of unknown function were observed to be significantly and reproducibly changing, and may lead to the annotation of additional candidates involved in Salt stress. The study also attempts to understand the general correlation of proteomics vs. transcriptomics data. Results from this work will lead to further studies with metabolic profiling, and gene deletion mutants.

* Presenting author

Oak Ridge National Laboratory and Pacific Northwest National Laboratory

8

Center for Molecular and Cellular Systems: High-Throughput Identification and Characterization of Protein Complexes

Michelle Buchanan[1], Frank Larimer[1], Steven Wiley[2], Steven Kennel[1], Dale Pelletier[1], Brian Hooker[2], Gregory Hurst[1], Robert Hettich[1], Hayes McDonald*[1] (mcdonaldwh@ornl.gov), Vladimir Kery[2], Mitchel Doktycz[1], Jenny Morrell[1], Bob Foote[1], Denise Schmoyer[1],Manesh Shah[1], and Bill Cannon[2]

[1]Oak Ridge National Laboratory, Oak Ridge, TN and [2]Pacific Northwest National Laboratory, Richland, WA

The Center for Molecular and Cellular Systems (CMCS) focuses on Goal 1 of the Genomics:GTL program, which aims to identify and characterize the complete set of protein complexes within a cell to provide a mechanistic basis of biochemical functions. Over the past two years, the CMCS has emphasized developing technologies that can be incorporated into a high throughput "pipeline" for the robust analysis of protein complexes. Several approaches for the isolation and identification of protein complexes from microbial cells were evaluated. Our experience has demonstrated that no single approach will be sufficient to handle the diverse types of complexes present in a cell. Thus, an integrated pipeline has been developed that uses two affinity-based approaches to isolate protein complexes in which tagged proteins are either expressed endogenously or exogenously. The individual technologies have been refined, validated and assembled into a semi-automated pipeline has been in operation for over 30 continuous weeks. A comprehensive laboratory information management system has been developed for sample tracking, process management, and data control. Experiments using over 200 tagged proteins have been conducted using *Rhodopseudomonas palustris* and *Shewanella oneidensis* cultures grown under different states. Data from the two types of pull-down approaches have been compared and have been found to provide complementary information. This suggests that both approaches are needed for comprehensive identification of protein complexes.

During the past year research tasks have been designed to improve the analysis pipeline. "Top-down" mass spectrometry has been used to identify modifications of constituent protein in the complexes. Several types of imaging tools have been employed to observe the complexes in live cells, including co-localization assays and fluorescence resonance energy transfer (FRET)-based assays. Additional effort has been placed on identifying new approaches for minimizing sample handling, such as microfluidic devices and automation. All of these research efforts have focused on development and validation of approaches to provide improved confidence of complex identification, increased sample throughput, and enhanced complex characterization.

* Presenting author

9

High-Throughput Analysis of Protein Complexes in the Center for Molecular and Cellular Systems

Vladimir Kery*[2] (vladimir.kery@pnl.gov), Dale A. Pelletier[1], Joshua N. Adkins[2], Deanna L. Auberry[2], Frank R. Collart[3], Linda J. Foote[1], Brian S. Hooker[2], Peter Hoyt[1], Gregory B. Hurst[1], Stephen J. Kennel[1], Trish K. Lankford[1], Chiann-Tso Lin[2], Eric A. Livesay[2], Tse-Yuan S. Lu[1], Cathy K. McKeown[1], Priscilla A. Moore[2], Ronald J. Moore[2], and Kristin D. Victry[2]

[1]Oak Ridge National Laboratory, Oak Ridge, TN; [2]Pacific Northwest National Laboratory, Richland, WA; and [3]Argonne National Laboratory, Argonne, IL

The Genomics:GTL Center for Molecular and Cellular Systems has implemented an integrated high-throughput "pipeline" for identifying the components of protein complexes from two bacterial species of interest to the DOE: *Shewanella oneidensis*, and *Rhodopseudomonas palustris*. This integrated pipeline uses two complementary approaches to isolate and identify protein complexes using affinity-tagged proteins—an endogenous approach, and an exogenous approach. In the exogenous approach, the targets of interest are cloned in a high-throughput procedure. Proteins are then expressed in *E. coli* and purified on Ni^{2+} agarose. Dialyzed purified tagged proteins are reattached to fresh Ni^{2+} agarose and exposed to lysate from the host cell of interest, thus forming protein complexes with host target proteins *in vitro*. In the endogenous approach, plasmids expressing the tagged protein of interest are transformed into the native host, and complexes are purified by tandem affinity purification using resins selective for the hexahistidine tag and the V5 epitope. In both approaches, the complexes are eluted from the beads under denaturing conditions, digested with trypsin and identified using automated liquid chromatography/electrospray tandem mass spectrometry in combination with SEQUEST™ analysis of the data. All main liquid handling procedures in protein and protein complex purification as well as MS sample preparation and MS measurement are automated. We are completing automation of data processing and bioinformatics. A laboratory information management system (LIMS) has been implemented for integrating all aspects of sample tracking, analysis and data flow. Between ORNL and PNNL, we have to date attempted expression of nearly 400 different genes as affinity-tagged fusion proteins, completed over 5000 "pulldown" experiments on these genes (including replicates) and identified several hundred different proteins in the pulldown samples. Distinguishing authentic interactors from non-specific interactors in the identified proteins has been an important aspect of this work. Our process was validated on a number of well known bacterial protein complexes; RNA polymerase, RNA degradosome, F1F0-ATP synthase, GroESL, and others. Some interesting findings on composition of other newly identified protein complexes are being further validated and investigated (e. g. peptidoglycan biosynthesis complexes involving genes Mur A, Mur C and Mur E of *S. oneidensis* etc.). While we have initially focused our efforts on *R. palustris* and *S. oneidensis*, the processes that we have developed are universally applicable to any organism of interest. Our aim is to scale up this process to provide a fully automated capability for high-throughput analysis of protein complexes with the goal of increasing throughput that would allow characterization of greater than 5,000 complexes per year.

* Presenting author

10

Investigating Gas Phase Dissociation Pathways of Crosslinked Peptides: Application to Protein Complex Determination

Sara P. Gaucher* (spgauch@sandia.gov), Masood Z. Hadi, and Malin M. Young

Sandia National Laboratories, Livermore, CA

Chemical crosslinking is an important tool for probing protein structure[1] and protein-protein interactions.[2-3] The approach usually involves crosslinking of specific amino acids within a folded protein or protein complex, enzymatic digestion of the crosslinked protein(s), and identification of the resulting crosslinked peptides by liquid chromatography/mass spectrometry (LC/MS). In this manner, distance constraints are obtained for residues that must be in close proximity to one another in the native structure or complex. As the complexity of the system under study increases, for example, a large multi-protein complex, simply measuring the mass of a crosslinked species will not always be sufficient to determine the identity of the crosslinked peptides. In such a case, tandem mass spectrometry (MS/MS) could provide the required information if the data can be properly interpreted. In MS/MS, a species of interest is isolated in the gas phase and allowed to undergo collision induced dissociation (CID). Because the gas-phase dissociation pathways of peptides have been well studied, methods are established for determining peptide sequence by MS/MS. However, although crosslinked peptides dissociate through some of the same pathways as isolated peptides, the additional dissociation pathways available to the former have not been studied in detail. Software such as MS2Assign[4] has been written to assist in the interpretation of MS/MS from crosslinked peptide species, but it would be greatly enhanced by a more thorough understanding of how these species dissociate. We are thus systematically investigating the dissociation pathways open to cross-linked peptide species. A series of polyalanine and polyglycine model peptides have been synthesized containing one or two lysine residues to generate defined inter- and intra-molecular crosslinked species, respectively. Each peptide contains 11 total residues, and one arginine residue is present at the carboxy terminus to mimic species generated by tryptic digestion. The peptides have been allowed to react with a series of commonly used crosslinkers such as DSS, DSG, and DST. The tandem mass spectra acquired for these crosslinked species are being examined as a function of crosslinker identity, site(s) of crosslinking, and precursor charge state. Results from these model studies and observations from actual experimental systems are being incorporated into the MS2Assign software to enhance our ability to effectively use chemical crosslinking in protein complex determination.

This project is funded by the DOE Genomics:GTL Program Project lead by ORNL/PNNL.

References

1. Young, MM; Tang, N; Hempel, JC; Oshiro, CM; Taylor, EW; Kuntz, ID; Gibson, BW; Dollinger, G. "High throughput protein fold identification by using experimental constraints derived from intramolecular crosslinks and mass spectrometry." *PNAS* 2000, *97*, 5802-5806.
2. Lanman, J; Lam, TT; Barnes, S; Sakalian, M; Emmett, MR; Marshall, AG; Prevelige, PE. "Identification of novel interactions in HIV-1 capsid protein assembly by high-resolution mass spectrometry." *J. Mol. Biol.* 2003, *325*, 759-772.
3. Rappsilber, J.; Siniossoglou, S; Hurt, EC; Mann, M. "A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry." *Anal. Chem.* 2000, *72*, 267-275.
4. Schilling, B; Row, RH; Gibson, BW; Guo, X; Young, MM. "MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides." *J. Am. Soc. Mass Spectrom.* 2003, *14*, 834-850.

* Presenting author

11

Center for Molecular and Cellular Systems: Statistical Screens for Datasets from High-Throughput Protein Pull-Down Assays

Frank W. Larimer*[1] (larimerfw@ornl.gov), Kenneth K. Anderson[2], Deanna L. Auberry[2], Don S. Daly[2], Vladimir Kery[2], Denise D. Schmoyer[1], Manesh B. Shah[1], and Amanda M. White[2]

[1]Oak Ridge National Laboratory, Oak Ridge, TN and [2]Pacific Northwest National Laboratory, Richland, WA

The large-scale analysis pipeline developed at ORNL and PNNL to identify protein complexes uses a high-throughput affinity-tag "pulldown" isolation step followed by denaturing elution, trypsin digestion and combined liquid chromatography tandem mass spectrometry analysis. Each pulldown experiment identifies potential associations between the target proteins and the tagged protein in the isolation step. Any protein complex analysis method may miss some protein:protein interactions and identify artifactual associations.

We are developing informatics-based criteria for assigning significance to protein identifications and associations. Data from blanks and quality assurance standards are used identify analysis problems such as carry-over between samples. Wild-type controls and replicate pull-downs are used to estimate repeatability. Proteins that show up with statistically significant frequency in a large number of experiments are used to establish background profiles. As our Mass Spec dataset of pulldown experiments grows, it will facilitate validation of this approach. With an understanding of the experimental *noise*, a quantitative estimate of the significance of *specific* pulldown results can be estimated.

We are also evaluating published statistical frameworks for interpreting protein association data. In tandem, we have developed a statistical screen for high-throughput pulldown experiments to reduce labeling spurious associations and strengthen identification of true associations. Initial results, though promising, emphasize the difficulties in developing a valid estimator of the probability of association between two proteins.

* Presenting author

12

Center for Molecular and Cellular Systems: Analysis and Visualization of Data from a High-Throughput Protein Complex Identification Pipeline Using Modular and Automated Tools

W. Hayes McDonald[1] (mcdonaldwh@ornl.gov), Joshua N. Adkins[2], Deanna L. Auberry[2], Kenneth J. Auberry[2], Gregory B. Hurst[1], Vladimir Kery[2], Frank W. Larimer[1], Manesh B. Shah[1], Denise D. Schmoyer[1], Eric F. Strittmatter[2], and Dave L. Wabb[1]

[1]Oak Ridge National Laboratory, Oak Ridge, TN and [2]Pacific Northwest National Laboratory, Richland, WA

Global or systems level analysis of biological processes is becoming increasingly common and some of the best examples are emerging out the field of proteomics. The Center for Molecular and Cellular systems is focused on high-throughput isolation and characterization of protein complexes. The core experimental pipeline of this effort uses the parallel and complementary approaches of affinity purification of endogenously expressed tagged proteins (endogenous pulldown) and heterologously expressed tagged proteins which are then used to isolate interacting proteins out of a cell lysate (exogenous pulldown). This integrated pipeline is currently being applied to the study of protein complexes from *R. palustris* and *S. oneidensis*. After isolation, constituents of these complexes are identified using high performance liquid chromatography coupled to either tandem mass spectrometry (LC-MS/MS) or high resolution mass spectrometry (LC-MS).

The two affinity isolation and the two mass spectrometry protocols have differing analysis requirements, therefore we have modularized our data analysis and visualization tools. This gives us not only the capabilities to automate and integrate data from these different sources, but also to "plug in" and evaluate new tools readily. Each of the following modules uses one or more software tools to accomplish its task: (a) MS data extraction and preparation - including extraction of data, filtering, and output to necessary format; (b) MS search – MS/MS database searches using SEQUEST or DBDigger or MS searches against an Accurate Mass Tag (AMT) database; (c) Search result filtering and summarization – protein identification and confidence; (d) Experimental filtering – reproducibility and background subtraction built on statistical evaluation and expert analysis; (e) Network visualization – using Cytoscape to view both simple and weighted networks of interactions. Currently, we have major modules automated; future work will require the seamless integration across modules and across PNNL and ORNL. Taken together this tool set gives us not only the ability to automate our data analysis, but also to quickly explore and compare relative strengths and utilities of both the experimental and data analysis pipelines.

Sandia National Laboratories

13

Carbon Sequestration in *Synechococcus*: A Computational Biology Approach to Relate the Genome to Ecosystem Response

Grant S. Heffelfinger* (gsheffe@sandia.gov)

Sandia National Laboratories, Albuquerque, NM

This talk will provide an update on the progress to date of the Genomics:GTL project led by Sandia National Laboratories: "Carbon Sequestration in *Synechococcus* sp.: From Molecular Machines to Hierarchical Modeling." This effort is focused on developing, prototyping, and applying new computational tools and methods to elucidate the biochemical mechanisms of the carbon sequestration of *Synechococcus* sp., an abundant marine cyanobacteria known to play an important role in the global carbon cycle. While much of our recent progress and results will be presented in detail in the seven or more posters submitted to this meeting (see Davidson et al., Geist et al., Martino et al., Plimpton et al., Samatova et al. Sinclair et al., Xu et al. and others), this talk will recap the larger focus and recent results of the project. Our project's results include experimental data on the *Synechococcus* carboxysome and CO_2 levels on growth rate and protein expression patterns in *Synechococcus* (sp. WH8102), the first characterizations of components of the proteome, and characterizations of the phosphorus and nitrogen regulatory pathways in conjunction with computationally derived predictions of these pathways. Our computational tool development efforts relative to processing high throughput experimental data have yielded new methods and algorithms for gene expression array analysis and a radically new tandem MS, MS/MS data analysis method which enables prototype assignment for large and diverse data sets (~60,000 spectra) with a surprising level of confidence. We have also developed and prototyped new computational tools for microbial systems biology, including methods for multi-scale characterization of protein interactions, methods for recognizing protein functional sites, and an integrating framework such tools. In addition to our work with *Synechococcus* Sp., we have applied these tools to other microbes in collaboration with other Genomics:GTL projects including the ORNL-PNNL microbial proteomics effort for *Rhodopseudomonas palustris* (with F. Larimer and H. McDonald). Our efforts to develop and apply modeling and simulation tools have yielded structural insight into the specificity of the carbon fixing enzyme RuBisCO as well as a computational capability to track spatial and temporal variations in protein species concentrations in realistic cellular geometries for important cyanobacterial subcellular processes. Finally, we have constructed an integrated data infrastructure which allows advanced search and queries across a large, diverse set of data sources (e.g. databases of sequence, structure, pathway, protein interaction, and raw mass spectra and microarray data). Our "*Synechococcus* Encyclopedia," contains all currently available database knowledge about this microbe and we are working now to create an encyclopedia for *Rhodopseudomonas palustris* and *Shewanella* for use by the GTL Microbial Complexes Pipeline project and *Shewanella* Federation respectively. More detailed discussions of our results may be found in our project's quarterly reports, available at www.genomes-to-life.org.

* Presenting author

14

Integrating Heterogeneous Databases and Tools for High Throughput Microbial Analysis

Nagiza Samatva* (samatovan@ornl.gov), Al Geist, Praveen Chandramohan, and Ramya Krishnamurthy

Oak Ridge National Laboratory, Oak Ridge, TN

Going beyond simple data archiving and retrieval of diverse data sets, we will describe a knowledge infrastructure that provides capabilities far beyond what has been available before. As part of the Genomics:GTL *Synechococcus*: From Molecular Machines to Hierarchical Modeling project, we have developed the technology needed to construct an integrated data infrastructure that allows advanced search and queries across a large, diverse set of data sources including sequence databases (COG, INTERPRO, SWISS PROT, TIGR, JGI, PFAM, PRODOM, SMART), structure databases (PDB, COILS, SOSUI, PROSPECT), pathway databases (KEGG), protein interaction databases (BIND, DIP, MIPS), and databases of raw mass spec and microarray data. Both a query language and integrated schema technology were developed to allow search and queries across these diverse databases.

We used our integrated data infrastructure to create a *Synechococcus* Encyclopedia (see Figure 1) containing all the database knowledge in the world about this microbe. This knowledgebase involves the integration of 23 different databases and is being used to do protein complex predictions, and pathway predictions. The technology can be used to create knowledgebases for other organisms and we have begun discussions with other GTL projects about setting up encyclopedias for their microbes.

The encyclopedia not only has data but also tools to analyze the data. This past year we have added a suite of easy-to-use web-based analysis tools to the encyclopedia. These tools, which are being developed within our GTL project, include protein function characterization, protein structure prediction, comparative analysis of protein-protein interfaces, metadata entry and browsing, pathway prediction, and electronic notebooks. Several of these tools provide transparent access to supercomputers at ORNL and around the nation. We will describe how the encyclopedia data and analysis tools were used to correctly predict the proteins making up a known membrane complex – including the membrane proteins involved – a task that is presently impossible by experimental mass spec analysis alone.

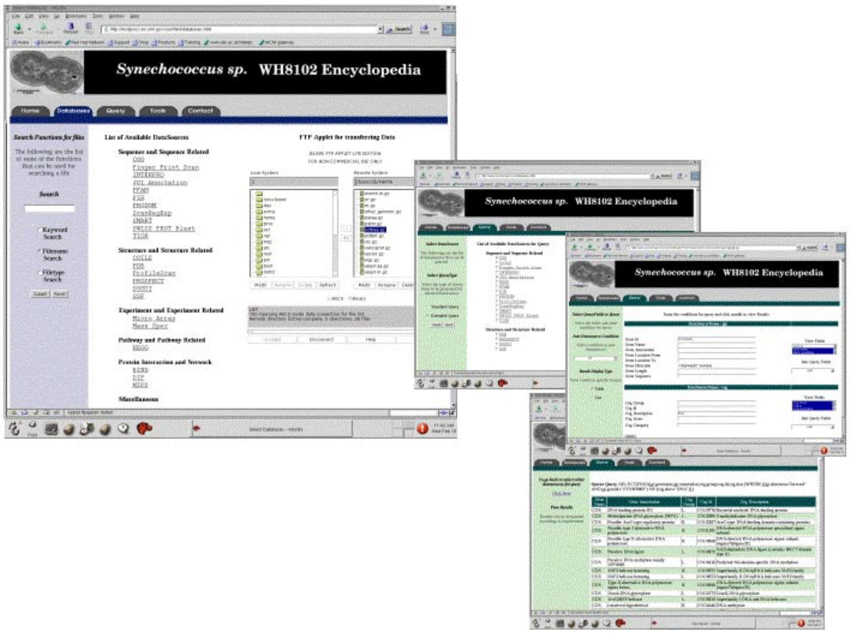
The new ability to rapidly construct advanced queries that require correlating and combining data from sequence annotations, protein structure, and interaction databases and to use the results in co-located analysis tools allows biologists to combine knowledge and see relationships that were previously obscured by the distributed nature and diverse data types in the biological databases.

The presentation will include "live" demonstrations of advanced queries of the *Synechococcus* Encyclopedia.

Acknowledgement: This project is supported by the U.S.Department of Energy's Genomics:GTL Program under project "Carbon Sequestration in *Synechococcus sp*: From Molecular Machines to Hierarchical Modeling" (http://www.genomes-to-life.org)

* Presenting author

Figure 1. *Synechococcus sp*. Encyclopedia. Advanced query and analysis interface. Search all *Synechococcus* databases. Browse experimental and analysis data. Download datasets.


Synechococcus sp. WH8102 Encyclopedia
Home
Databases
Query
Tools
Contact
Synechococcus sp. WH8102 Encyclopedia
Synechococcus sp. WH8102 Encyclopedia

15

Toward Comprehensive Analysis of MS/MS Data Flows

Andrey Gorin* (agor@ornl.gov), Nikita D. Arnold, Robert M. Day, and Tema Fridman

Oak Ridge National Laboratory, Oak Ridge, TN

Tandem mass spectrometry (MS/MS) is a powerful tool applied across several Genomics:GTL projects for a variety of challenging proteomics projects: search for modified proteins, characterization of whole cell proteome, and identification of components of protein molecular machines. Despite great variety of the biological drivers, computational algorithms used "under the hood" face exactly the same challenges, and existing limitations of such algorithms are reproduced across many experimental designs. in ion trap devices under common conditions only ~20% of MS/MS spectra lead to peptide identifications that are worth to be considered, and misidentification rates remains to be high.

In certain range of score values the problem presents the tug-of-war alternative — boost of the reliability threshold (e.g. SEQUEST x-correlation value) rapidly decreases fraction of spectra that could be identified, while lowering it produces identifications from the "grey area", which are of dubious quality. Algorithmically, the only way out is to increase *information extraction* from tandem MS data. If we could somehow retrieve total information content of a given spectrum, its fate can be decided unambiguously depending on our capacity to learn from it. Such capability could be useful for a

* Presenting author

number of other interesting proteomics applications. The difficulties, of course, start with the definition of something as unusual as information content of peptide spectrum.

Recently we proposed Probability Profile Method (PPM) — classification algorithm that infers identities of the individual spectral peaks examining their spectral neighborhoods under the "microscope" of Bayesian statistics. PPM results have the form of probabilistic statements, like *peak number 123 is a b-ion with a 0.85 probability*. Efficient identification of "noble" b- and y-ion peaks dramatically simplifies construction of *de novo* tags (partial peptides) for a particular spectrum. Relatively simple algorithmic advances allowed us to build PPM-chain – tool for de novo protein tagging based on our methodology. During this study we have realized that traditional separation of MS computational algorithms into database search and de novo is very misleading. Our PPM-chain can be used in SEQUEST-emulation mode, taking full advantage of the known protein database, but at the same time has quite unique algorithmic capabilities, which include classical full-length de novo sequencing (it is not very good at the later task yet).

While capable of emulating SEQUEST our program works on entirely different mathematical and algorithmic principles. The laborious comparison between theoretical spectra and experimental spectra is the main CPU time consumer in database look-up algorithms, and correspondingly the performance typically scales linearly with the size of the search space, which grows exponentially in many situations (e.g., with the number of PTMs considered for each peptide). In de novo approach, almost all work is done up front, on the experimental spectra: peak labeling, finding of the tags, tag scoring. The need for the database comes very late in the process, involves very few candidate sequences and very simple procedures, which could be skipped all together for spectra with too little (no connectable peaks) or a lot (direct de novo identification) in terms of the informational content.

De novo identification also has inherent flexibility, which is reflected in the suppleness of its output. For a given spectrum and given specifications for de novo tag (e.g. 3 residues are set as a minimum length) PPM-chain has three possible outcomes: (1) "no tag" - no satisfactory de novo tag could be constructed for the spectrum; (2) "tag-no-match" – there are good de novo tags, but they do not conform to the available database; (3) "answer" – satisfactory de novo tag is found and mapped to a protein in the database. In contrast, database look-up programs return the best match with an attached score, which slowly decreases from the confidently identified spectra toward definite identification failures. In this case the bad quality of the match (e.g., because the database protein contains sequencing error) is hard to distinguish from the mediocre informational content of the spectrum (e.g., due to poor fragmentation). Such mix-up leads to all kinds of "grey area" situations, where valuable information - often indicative of unusual and interesting biological events - can be irrecoverably lost

We compared PPM-chain and SEQUEST using data sample obtained on the 54 ribosomal proteins of the *Rhodopseudomonas palustris* produced by Dr. Michael Strader at ORNL Center for Molecular and Cellular System. We have explored results of both programs for three spectral sets separated by SEQUEST X-correlation score: "high quality" (>3.2), "medium" (from 2.2 to 3.2) and "low quality" (<2.2) spectra. For the high quality subset "no tag" outcome was obtained only for 21 spectra (1.4%) and out of 1263 "answer" results SEQUEST identification was confirmed for 99.9% spectra. "Tag-no-match" outcome was observed for 216 cases (14%) and this fraction kept increasing in medium and low quality subsets: (~ 38% in both). The "answer" outcomes were still excellently aligned with SEQUEST ids (99% and 96% precision values, correspondingly). The fraction of "no tag" cases has grown sharply: 18% for medium and 57% for low confidence sets, reflecting the absence of the differentiating information in many spectra belonging there.

The result suggests an interesting speculation about possible sources of SEQUEST errors: it is feasible that a large fraction of such errors is due to the absence of the underlying correct answer in the database. In such cases the returned match bound to be an incorrect one, but still may have relatively high X-correlation value. In our approach such spectra immediately become candidates for further study, such as Post Translational Modification (PTM) search or further de novo processing.

Summarizing, our testing indicates the following:

- Even with the existing technology (which certainly could be improved) reliable *de novo* tags can be constructed for a large majority of MS/MS spectra – and virtually for all high quality spectra.
- When de novo solution is compatible with the database, it is almost always the same as provided by SEQUEST. This conclusion confirms a high reliability of the SEQUEST identifications in the cases when the expected peptides are present in the protein database.
- There is a significant fraction of spectra (~33% for medium X-correlation values, ~50% low X-corr values), where the PPM-Chain finds good de novo tags not compatible with anything in the target database. Some of these tags definitely reflect complex and interesting biological phenomena, where PTMs and point mutations are blocking the possibility of finding the correct answers in the "plain vanilla" database searches.

For a future work we plan to apply PPM-chain for a comprehensive data extraction from the proteomics samples aiming at low abundance proteins as well as interesting biological facts, such as PTMs and mutated proteins. Our results strongly suggest that this approach will not only increase the output of useful information, but will also eliminate significant part of incorrect identification, further improving quality of the corresponding proteomics studies.

This work was funded in part by the US Department of Energy's Genomics:GTL program (www.doegenomestolife.org) under two projects, "Carbon Sequestration in *Synechococcus Sp.*: From Molecular Machines to Hierarchical Modeling" (www.genomes-to-life.org) and "Center for Molecular and Cellular Systems" (www.ornl.gov/GenomestoLife).

16

The Transcriptome of a Marine Cyanobacterium—Analysis Through Whole Genome Microarray Analyses

Brian Palenik[1]* (bpalenik@ucsd.edu), Ian Paulsen[2]* (ipaulsen@tigr.org), Bianca Brahamsha[1], Rob Herman[1], Katherine Kang[2], Ed Thomas[3], Jeri Timlin[3], and Dave Haaland[3]

[1]Scripps Institution of Oceanography, La Jolla, CA; [2]The Institute for Genomic Research, Rockville, MD; and [3]Sandia National Laboratories, Albuquerque, NM

Nitrogen and phosphorus abundance and type are thought to control photosynthesis and carbon sequestration in large areas of the world's oceans. Little is known about the regulation in cyanobacteria of nitrogen and phosphorus metabolism and their interaction with other environmental variables such as light and micronutrients. We are using a whole genome microarray of *Synechococcus* sp. WH8102 to examine these issues.

We have used whole genome microarrays in a number or experiments, initially to compare cells grown with nitrate and cells grown with ammonia. We found that 247 genes were down-regulated

* Presenting author

during growth under ammonia compared to nitrate. This included the NtcA transcriptional regulator known to control growth when ammonia is low and nitrogen sources other than ammonia are used. We also found that the use of a number of alternative nitrogen sources were down regulated (e.g. nitrate metabolism, cyanate transport, and urea metabolism). Some of these clearly have ntcA binding sites upstream although a complete comparison of the microarray results with ntcA binding site predictions are still in progress by Zhengchang Su and Ying Xu (UGA) as part of our GTL project. In addition, a number of genes associated with stress conditions are down regulated. These include glutathione peroxidase and a number of proteases and heat shock like proteins. This supports our hypothesis that growth under nitrate is actually more stressful than on ammonia due to the requirement of additional electron transport activity (and electron leakage) to reduce nitrate to nitrite to ammonia. Thus, we currently hope to analyze and model these results as a combination of NtcA regulation and stress response regulation. Interestingly a number of hypotheticals and conserved hypotheticals are down regulated, giving us an initial clue as to their possible functions in the cell.

In addition, we have also characterized phosphate limitation in WH8102 and made knockout mutants in a number of the two-component regulatory systems of the cell. We are also examining these phosphate limitation experiments with the wild type and mutant cells using whole genome microarray analyses. Because of its relatively small number of regulatory systems compared to many microbes, *Synechococcus* sp. WH8102 is an ideal model system for preparing a complete picture of the regulatory networks of an environmentally significant microbe.

Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-ACO4-94AL85000.

This work was funded by the U.S.Department of Energy's Genomics: GTL program (www.doegenomestolife.org) under project, "Carbon Sequestration in *Synechococcus Sp.*: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org) and by its Microbial Genome Project under "Transport and its regulation in marine microorganisms—a genomics-based approach."

17

DEB: a Data Entry and Browsing Tool for Entering and Linking *Synechococcus sp.* WH8102 Whole Genome Microarray Metadata from Multiple Data Sources

Arie Shoshani[1]* (Shoshani@lbl.gov), Victor Havin[1], Vijaya Natarajan[1], Tony Martino[2], Jerilyn A. Timlin[2], Katherine Kang[3], Ian Paulsen[3], Brian Palenik[4], and Thomas Naughton[5]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]Sandia National Laboratories, Albuquerque, NM; [3]The Institute for Genomic Research, Rockville, MD; [4]Scripps Institution of Oceanography, La Jolla, CA; and [5]Oak Ridge National Laboratory, Oak Ridge, TN

The process of generating and analyzing microarray data for *Synechococcus sp.* WH8102 whole genome in the Sandia-led GTL project involves three collaborators, where each generates metadata about their operation as well as data files. The *Synechococcus sp.* microbes are cultured in the Scripps Institution of Oceanography in San Diego, then the sample pool is sent to TIGR in Rockville, Maryland for microarray hybridization, 2-color scanning, and analysis. The scanned files and slides are then sent to Sandia Lab in Albuquerque, New Mexico for analysis and additional scanning with

* Presenting author

a Hyperspectral Imaging instrument. Each of the institutes has an independent system for keeping track of metadata about their part of the operation and unfortunately these systems do not facilitate easy transfer of metadata details between institutions. This situation is typical of many biology projects, and it begs for a solution.

In this sub-project we set out to develop a single system where such metadata can be collected and linked in an orderly fashion. We developed a web-based Data Entry and Browsing (DEB) tool that can capture the metadata from experiments and laboratories and store them in a database in a computer searchable form. The key need is to have an easy-to-use intuitive system that integrates the metadata of all the related activities in this project. The design of the DEB tool is based on inputs and insights from the biologists on the project and as such contains features that a biologist will find useful. The interface design mimics the familiar laboratory notebook format. The system is built on top of the Oracle database system. The main concept of the interface design is to expose the biologist to a single "object" and its attributes at a time, and presenting objects as pages in a notebook that can be "turned", yet provide links between the objects in a simple intuitive fashion. An example of such a web-based screen is shown in the figure below.

The most powerful capability of the DEB system is that it is schema-driven, that is, all the interfaces to support all of the above features are generated automatically from the schema definition. Therefore, new metadata schemas can quickly be used to generate DEB interfaces as well as the underlying Oracle database for them. This feature makes this tool immediately applicable to new and/or changing databases. This allowed us to generate databases based on schemas designed by the biologists. Specifically, the scientists from the three sites have defined schemas for the Nucleotide Pool of microbes, for the Microarray Hybridization (based on the MIAME concepts), and the Hyperspectral Imaging and analysis system. The design included the ability to link these schemas and thus allow a researcher from any area of the project to extract metadata from the various parts of the experiment. For example the microarray hybridization schema has "probe_source" that links (points to) the "nucleotide_pool_id" in the Nucleotide Pool schema.

Data entry to the databases is done in two different modes: 1) on-line web-based data entry, and 2) automated data uploading from another database source. The on-line mode is used by the people who culture the Nucleotide Pools (Scripps), and the people running the Hyperspectral Imaging (Sandia). The automated data uploading is used for the Microarray Hybridization metadata (TIGR) because they have their own well-developed internal database system. The automated metadata loading is performed by dumping the metadata into simple formatted files (similar the spreadsheet output format) and have schema-driven tools for loading the data into the common database.

The main features of the DEB system are:

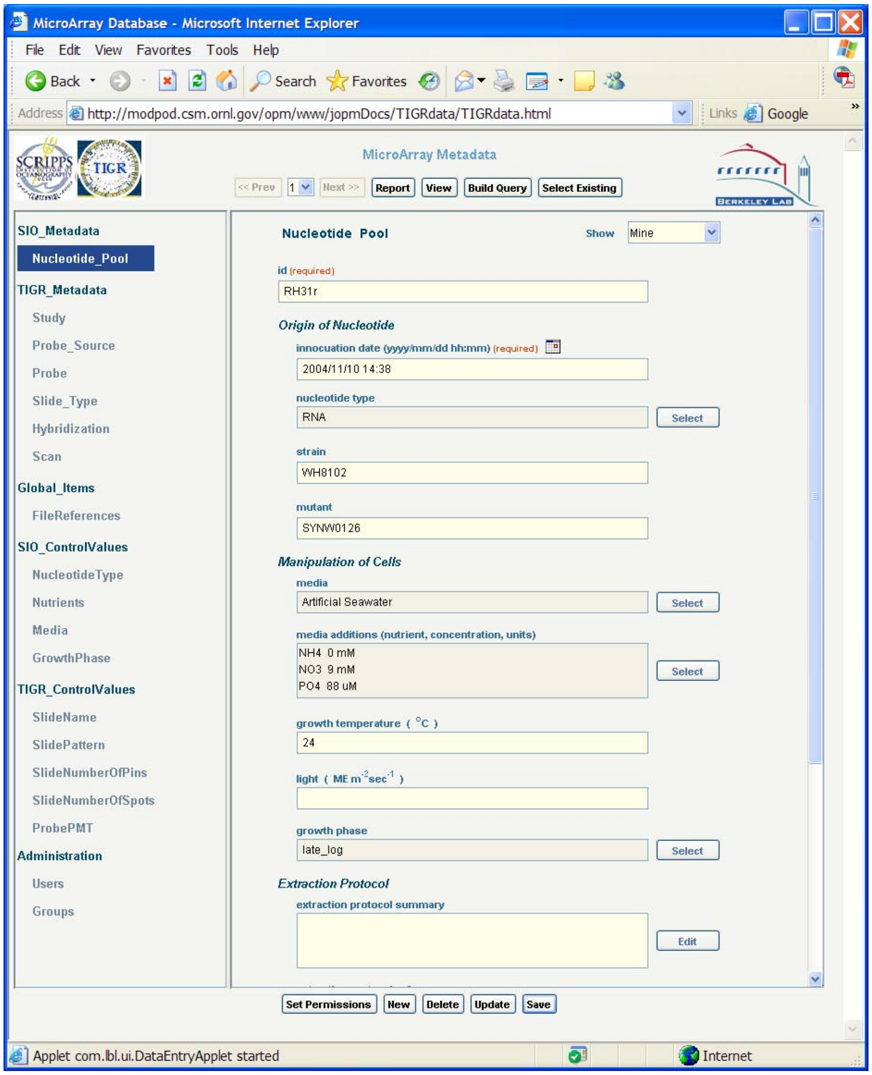
- It supports multiple inter-related object-classes, such as "experiment, materials, nucleotide-pool, samples, arrays, etc.
- For each object-class, it displays a page that mimics a notebook, with pages that can be "turned" (i.e. selected by previous-next, or by number)
- Objects can be linked to each other by simple connectors, such as a "sample" object linked to its "nucleotide-pool".
- Any file types (document, images, excel, etc.) can be uploaded to the system, and related to the metadata

* Presenting author

- Pages of the metadata can be printed for entry into a physical notebook – a requirement that makes sure the information is physically recorded
- Recording a new entry can be based on a previous entry, thus avoiding the re-entry of existing entries
- Security features to protect the metadata can be controlled by users and groups; each experiment or related objects can be assigned read, write, and delete permission to other users/groups.
- A query feature to search the metadata based on conditions on the attributes of the objects.

The importance of an easy-to-use system for capturing metadata in GTL cannot be overlooked, especially as an ever growing number of experiments are conducted and a large number of datasets are collected. The ability to quickly and automatically generate metadata systems from a schema description is essential for this evolving field with multiple sources of data gathered independently. DEB is currently running on LBNL's development server and at ORNL's GTL project operational server. While this system is designed for this project, its schema-driven architecture means that it can be applied to other GTL efforts.

Sandia is a multi-program laboratory operated by Sandia Corpora-



MicroArray Database - Microsoft Internet Explorer

Address http://modpod.csm.ornl.gov/opm/www/jopmDocs/TIGRdata/TIGRdata.html

MicroArray Metadata

Nucleotide Pool

Show: Mine

id (required): RH31r

Origin of Nucleotide

innocuation date (yyyy/mm/dd hh:mm) (required): 2004/11/10 14:38

nucleotide type: RNA

strain: WH8102

mutant: SYNW0126

Manipulation of Cells

media: Artificial Seawater

media additions (nutrient, concentration, units):
NH4 0 mM
NO3 9 mM
PO4 88 uM

growth temperature (°C): 24

light (ME m^{-2}sec^{-1}):

growth phase: late_log

Extraction Protocol

extraction protocol summary:

Applet com.lbl.ui.DataEntryApplet started

* Presenting author

tion, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-ACO4-94AL85000.

18

Microarray Analysis using VxInsight and PAM

George S. Davidson*[1] (GSDAVID@sandia.gov), David Hanson[2], Shawn Martin[1], Margaret Werner-Washburne[2], and Mark D. Rintoul[1]

[1]Sandia National Laboratories, Albuquerque, NM and [2]University of New Mexico, Albuquerque, NM

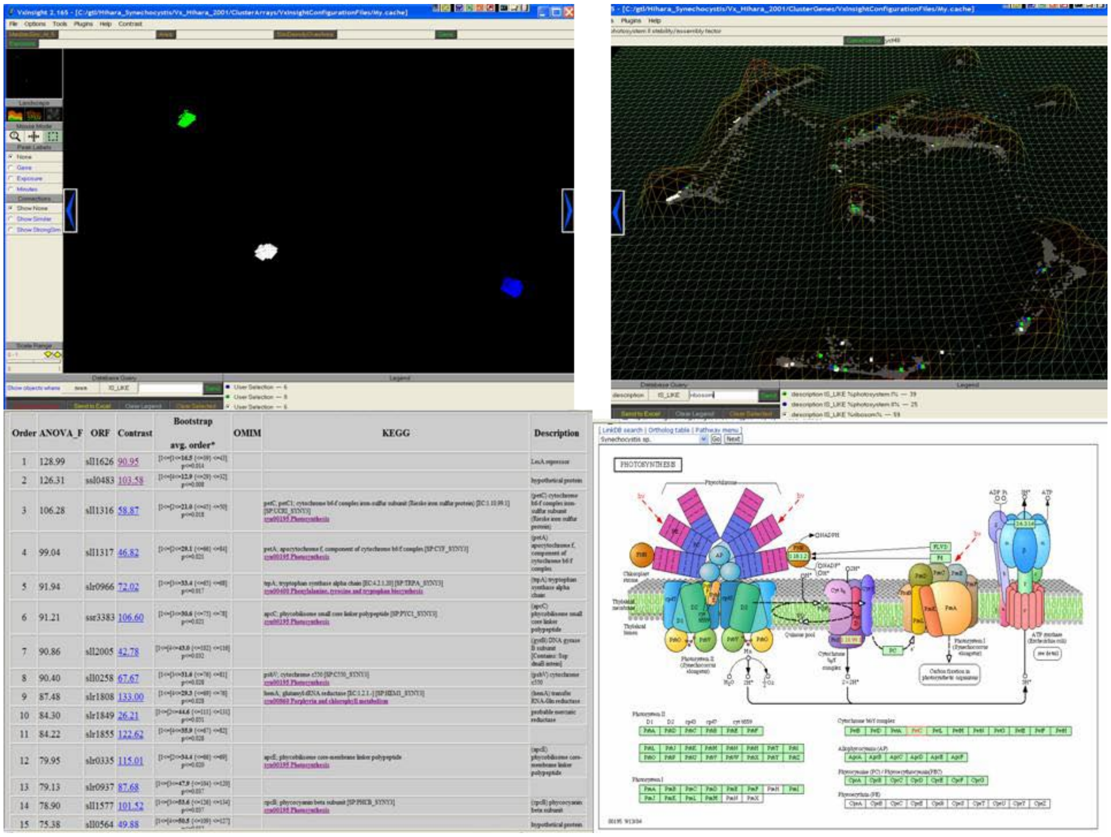
In 2001 Hihara *et al.* [1] published a series of microarray experiments describing gene expression changes in the cyanobacterium *Synechocystis sp.* PCC 6803 in response to acclimation from a low light level (20 µmol photons m^{-2} sec^{-1} to 300 µmol photons m^{-2} sec^{-1}). These data, which are publicly available from the KEGG database [2], have been reanalyzed using the VxInsight genome tools [3] from Sandia National Laboratories, and PAM [4] from Stanford. The analysis served as a test-bed for preparing the VxInsight tools to work with bacterial microarray data and associated databases. Here we present the results of clustering the individual experiments and the genes by co-expression. Interestingly, a number of experimental design issues are also raised. We examined lists of genes generated by PAM and by VxInsight that include significant differences in expression under the low light (LL) and the high light (HL) experimental conditions. We discuss the methodologies and the visual user interface linking these genes to online annotations and regulatory networks.

Hihara *et al.* measured total mRNA from HL conditions at 15 min, 1 hr, 6 hr, and 15 hr. These were compared to the mRNA levels measured under LL conditions, which served as control data. Expression levels were measured with CyanoCHIP version 0.8 from TaKaRa. These experiments revealed 84 ORFs with up regulated expression and 80 ORFs with down regulated expression after exposure to HL. Almost all of the photosystem I genes were immediately down regulated, while genes associated with photosystem II showed more complicated patterns. Both observations are consistent with the increasing PSII/PSI ratio in response to HL which involves an initial shift away from of PSI and the gradual construction of PSII (generally completed within 60 min.).

VxInsight found three groups of experiments, as shown in Figure 1, the first of which clearly reflects the stable, ongoing response to HL. The second captures the intermediate state when many of the PSII proteins are being synthesized, or have largely become available as the cells shift toward a higher metabolic plane. The third group contains a random mixture of arrays from time points throughout the experiment, which suggests that technical problems were confounding the measurements (something that is not uncommon in microarray experiments). We identified genes with significantly different expressions between late experimental conditions (first group) and the second group, which consisted of an equal number of measurements made at 15 minutes and at 60 minutes. VxInsight and PAM identified many of the same genes that Hihara *et al.* found; however the differences offer opportunities for deeper study. We present these genes with their scores and demonstrate the interactive analysis of these lists, including the use of KEGG pathways.

* Presenting author

Figure 1. The three clusters of arrays in the Hihara et al. experiment (top left), together with the gene clusters (top right). The top 15 genes relevant to the shift from HL to LL are listed on the bottom left, where the list contains links to KEGG networks, as shown on the bottom right.



Order	ANOVA_F	ORF	Contrast	Bootstrap avg. order*	OMIM	KEGG	Description
1	128.99	sll1626	90.95	[1<=[1<=14.5 (<=19) <=43] p<=0.014			LexA repressor
2	126.31	ssl0483	103.58	[1<=[4<=12.9 (<=29) <=32] p<=0.008			hypothetical protein
3	106.28	sll1316	58.87	[1<=[2<=21.0 (<=43) <=50] p<=0.018		petC, petC1; cytochrome b6-f complex iron-sulfur subunit (Rieske iron sulfur protein) [EC:1.10.99.1] [SP:UCRI_SYNY3] syn00195 Photosynthesis	(petC) cytochrome b6-f complex iron-sulfur subunit (Rieske iron sulfur protein)
4	99.04	sll1317	46.82	[1<=[2<=29.1 (<=66) <=84] p<=0.021		petA; apocytochrome f, component of cytochrome b6-f complex [SP:CYF_SYNY3] syn00195 Photosynthesis	(petA) apocytochrome f, component of cytochrome b6-f complex
5	91.94	slr0966	72.02	[1<=[1<=33.4 (<=65) <=88] p<=0.017		trpA; tryptophan synthase alpha chain [EC:4.2.1.20] [SP:TRPA_SYNY3] syn00400 Phenylalanine, tyrosine and tryptophan biosynthesis	(trpA) tryptophan synthase alpha chain
6	91.21	ssr3383	106.60	[1<=[1<=30.6 (<=73) <=78] p<=0.021		apcC; phycobilisome small core linker polypeptide [SP:PYC1_SYNY3] syn00195 Photosynthesis	(apcC) phycobilisome small core linker polypeptide
7	90.86	sll2005	42.78	[1<=[4<=43.0 (<=102) <=116] p<=0.032			(gyrB) DNA gyrase B subunit [Contains: Ssp dnaB intein]
8	90.40	sll0258	67.67	[1<=[1<=31.6 (<=78) <=81] p<=0.028		psbV; cytochrome c550 [SP:C550_SYNY3] syn00195 Photosynthesis	(psbV) cytochrome c550
9	87.48	slr1808	133.00	[1<=[4<=29.3 (<=69) <=78] p<=0.028		hemA; glutamyl-tRNA reductase [EC:1.2.1.-] [SP:HEM1_SYNY3] syn00860 Porphyrin and chlorophyll metabolism	(hemA) transfer RNA-Glu reductase
10	84.30	slr1849	26.21	[1<=[2<=44.6 (<=111) <=131] p<=0.031			probable mercuric reductase
11	84.22	slr1855	122.62	[1<=[4<=35.9 (<=67) <=82] p<=0.028			
12	79.95	slr0335	115.01	[1<=[2<=34.4 (<=66) <=69] p<=0.020		apcE; phycobilisome core-membrane linker polypeptide syn00195 Photosynthesis	(apcE) phycobilisome core-membrane linker polypeptide
13	79.13	slr0937	87.68	[1<=[3<=47.9 (<=104) <=120] p<=0.037			
14	78.90	sll1577	101.52	[1<=[3<=53.6 (<=126) <=134] p<=0.037		cpcB; phycocyanin beta subunit [SP:PHCB_SYNY3] syn00195 Photosynthesis	(cpcB) phycocyanin beta subunit
15	75.38	sll0564	49.88	[1<=[4<=50.5 (<=108) <=127]			hypothetical protein

References

1. Hihara, Y., et al., *DNA microarray analysis of cyanobacterial gene expression during acclimation to high light.* The Plant Cell, 2001. **13**: p. 793-806.

2. Hihara, Y., http://www.genome.jp/kegg/expression/.

3. Davidson, G.S., *et al.*, *High throughput instruments, methods, and informatics for systems biology* (also to appear as: *Robust Methods for Microarray Analysis,* in *Genomics and Proteomics Engineering in Medicine and Biology*, IEEE/Wiley Press, Metin Akay, editor, in press*). 2003, Sandia National Laboratories SAND2003-4664: Albuquerque, New Mexico 87185.*

4. *Tibshirani, R., et al., Diagnosis of multiple cancer types by shrunken centroids of gene expression.* PNAS, 2002. **99**(10): p. 6567-6572.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. This work was supported by the U.S.Department of Energy's Genomics: GTL program (www.doegenomesto-life.org) under project, "Carbon Sequestration in *Synechococcus sp.*: From Molecular Machines to Hierarchical Modeling,"(www.genomes-to-life.org).

* Presenting author

19

Mapping of Biological Pathways and Networks across Microbial Genomes

F. Mao, V. Olman, Z. Su, P. Dam, and Ying Xu* (xyn@bmb.uga.edu)

University of Georgia, Athens, GA and Oak Ridge National Laboratory, Oak Ridge, TN

Homology exists beyond the individual gene level, and it could exist at the biological pathway and network level. There are a number of databases consisting of all experimentally validated and reliably predicted pathways/networks, providing a rich source of information for genome annotation and biological studies at a systems level. A key to effectively use such information is to identify orthologous genes accurately. However existing methods for mapping these known pathways and networks have serious limitations, greatly limiting the utility of such very useful information. Virtually all existing mapping methods are based on sequence similarity information, using tools such as reciprocal BLAST search or COG mapping. A fundamental problem with such methods is that sequence similarity information alone does NOT contain all the information needed to identify true orthologous genes!

We have recently developed a computational method and software, called P-MAP, for mapping a known pathway/network from one microbial organism to another by combining homology information and genomic structure information. The basic idea is that in microbes, genes working in the same pathway can generally be decomposed into a few operons or, in case of complex pathways/networks, regulons. Such information has not been effectively used in pathway mapping. When mapping known pathways, we first predict all the operons in a genome using our operon prediction program. The predictions are then validated through comparing microarray data mainly to check for consistency between gene expression patterns for genes predicted to be in the same operons or adjacent operons. Our evaluation has indicated that our prediction accuracy is close to 90%. With such information, we then map genes in a pathway template to the target genome that simultaneously gives relatively high sequence similarity between predicted orthologous gene pairs and has all the mapped genes grouped into a number of operons, preferably co-regulated operons based on the predicted *cis* regulatory elements and available microarray data. We have formulated the mapping problem as a linear integer programming (LIP) problem, and solved the problem using a commercial LIP solver, called COIN.

We have applied the P-MAP program to map known biological pathways in KEGG and MetaCyc to the cyanobacterial genomes and currently are mapping them to the *Shewanella oneidensis* MR-1 genome. Some of the mapping results could be found at http://csbl.bmb.uga.eddu/WH8102.

Acknowledgement: This project is supported by the U.S.Department of Energy's Genomics:GTL Program under project "Carbon Sequestration in *Synechococcus sp*: From Molecular Machines to Hierarchical Modeling" (http://www.genomes-to-life.org).

* Presenting author

20

Proteomic Analysis of the *Synechococcus* WH8102 CCM with Varying CO_2 Concentrations

Arlene Gonzales, Yooli K. Light, Zhaoduo Zhang, Michael D. Leavell, Rajat Sapra, Tahera Iqbal, Todd W. Lane, and Anthony Martino* (martino@sandia.gov)

Sandia National Laboratories, Livermore, CA

The genera *Synechococcus* and *Prochlorococcus* are oxygenic photoautotroph cyanobacteria. They are the most abundant picophytoplankton in the world's oceans where they form the foundation of the marine food web and are likely the largest contributors to primary production. Whole genome sequences are now available for a number of cyanobacteria including *Synechococcus* WH8102, *Prochlorococcus* MED4, and *Prochlorococcus* MIT9313. The sequences make it possible to use comparative analysis and high-throughput functional genomics and proteomics experiments to help better understand global diversity involved in carbon fixation.

Synechococcus WH8102's 2.4 Mb genome has yielded a number of interesting results regarding the carbon concentrating mechanism (CCM) in this organism. The carboxysome encoding operon in 8102 resembles that of β-proteobacteria rather than cyanobacteria. The operon most likely was acquired through horizontal gene transfer from phage. Carbonic anhydrase (CA) activity in the carboxysome shell protein csoS3 has been determined experimentally. Genome analysis indicates a putative β-CA and a ferripyochelin binding protein CA may also exist. Finally, transport of inorganic carbon in 8102 may occur through the low affinity CO_2 uptake genes ndhD4, ndhF4, and chpX. In *Prochlorococcus*, uptake genes have not been observed. Perhaps a unique transport mechanism exists in oceanic cyanobacteria.

We will present a high-throughput proteomic approach using mass spectrometry (MS), 2-hybrid analysis, and phage display to deconstruct components of the CCM and determine the effect of changing CO_2 levels in *Synechococcus* 8102. Protein expression levels of CCM components and protein-protein interactions within the carboxysome will be presented. Protein fractions were separated in to particulate and soluble fractions, and western blots of the fractions indicated rbcL and carboxysome shell proteins partitioned exclusively with the particulate fractions. Developing and fully mature carboxysomes were observed in the particulate fractions using electron microscopy. The two putative CAs partitioned separately in the particulate and soluble fractions. Changes in expression of specific proteins in cultures bubbled under different CO_2 levels were determined using 2D electrophoresis/MALDI-TOF MS. A synergistic whole proteome approach using capillary LC-MS/MS continues. Protein-protein interactions within the carboxysome have been determined using bacterial 2-hybrid techniques, and a number of pair wise interactions with be presented. Finally, rbcS-peptide interactions are being studied using phage display techniques.

* Presenting author

21

Predicting Protein-Protein Interactions Using Signature Products with an Application to β-Strand Ordering

Shawn Martin[1] (smartin@sandia.gov), W. Michael Brown[1], Charlie Strauss[2], Mark D. Rintoul*[1], and Jean-Loup Faulon[3]

[1]Sandia National Laboratories, Albuquerque, NM; [2]Los Alamos National Laboratory, Los Alamos, NM; and [3]Sandia National Laboratories, Livermore, CA

As a part of the project entitled "Carbon Sequestration in *Synechococcus sp.*: From Molecular Machines to Hierarchical Modeling," we have developed a computational method for predicting protein-protein interactions from amino acid sequence and experimental data (Martin, Roe et al. 2004). This method is based on the use of symmetric tensor products of amino acid sequence fragments, which we call *signature products*. These products occur with different frequencies when considering interacting versus non-interacting protein pairs. We can therefore predict when a protein pair interacts by comparing the frequency of the signature products in that pair to the corresponding frequencies in protein pairs known to interact. Computationally, these comparisons are encoded into a Support Vector Machine (SVM) framework (Cristianini and Shawe-Taylor 2000), where the signature products are implemented using kernel functions. The final result is an automated classification system which can extrapolate from experimental results to complexes to entire proteomes, based only on experiment and primary sequence.

We have expended significant effort in benchmarking our method against competing techniques. In particular, we have compared the signature product method with methods based on products of InterPro signatures (Sprinzak and Margalit 2001; Mulder, Apweiler et al. 2003), concatenation of full-length amino acid sequences (Bock and Gough 2001), and a method combining multiple data sources (Jansen, Yu et al. 2003). In all cases our method performed as well as or better than the competing methods, as measured by 10-fold cross validation. We have applied our method to the prediction of protein-protein interactions in the case of yeast SH3 domains (Tong, Drees et al. 2002), where we achieved 80.7% accuracy, the full yeast proteome (69% accuracy), and *H. Pylori* (Rain, Selig et al. 2001) (83.4% accuracy). In addition to these results, which appear in (Martin, Roe et al. 2004), we have applied our method using COG networks (Tatusov, Koonin et al. 1997) to *Synechocystis sp.* (91% accuracy), and *Nostoc sp.* (69% accuracy).

Using our signature product approach, we have gone on to develop a method for ordering β-strands, which can in turn be used to improve the results of *ab initio* protein folding. The first step in our method is to train a signature product model for predicting β-strand interactions. This model was trained by extracting all β-strands from the Protein Data Bank (PDB) (Berman, Westbrook et al. 2000) using the dictionary of protein secondary structure (DPSS) method (Kabsch and Sander, 1983). After removing identical sequences from the ~20,000 structures available in the PDB, we obtained ~90,000 β-strands with more than 3 residues. Using the product signature method, we trained a β-strand interaction predictor which achieved 77% accuracy on a randomly selected training/test set combination with 80% of the β-strands in the training set and the remaining 20% of the β-strands in the test set. The next step in our method, as yet unimplemented, applies to individual proteins. For a given protein, we will apply our model to every possible pair of β-strands within the protein, and then consider every possible ordering of these strands. We will use the ordering which gives the highest

* Presenting author

average interaction score, as measured using our β-strand interaction predictor. We will validate our results by comparing our predicted ordering to the ordering actually present in the PDB.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

22

In Vivo Observation of the Native Pigments in *Synechocystis sp.* PCC 6803 Using a New Hyperspectral Confocal Microscope

Michael B. Sinclair[1]* (mbsincl@sandia.gov), Jerilyn A. Timlin[1], David M. Haaland[1], Sawsan Hamad[2], and Wim F.J. Vermaas[2]

[1]Sandia National Laboratories, Albuquerque, NM and [2]Arizona State University, Tempe, AZ

We have developed a new hyperspectral confocal microscope that combines the attributes of high spatial resolution (<0.5 μm), high speed acquisition (>8 MB/s) and single photon sensitivity. The new instrument records the emission spectrum from 500 nm to 800 nm for each voxel within the 3-dimensional sample. The acquisition of full spectral information, when coupled with modern multivariate data analysis techniques allows for quantification of the contribution of each of the emitting components present within any voxel. To demonstrate the advantages of this approach, we have obtained and analyzed in vivo hyperspectral images of wild type *Synechocystis sp*. PCC 6803, as well as two mutant strains: chlL–, which is incapable of light-independent synthesis of chlorophyll, and PS1-less/chlL–, which in addition to being incapable of light-independent synthesis of chlorophyll does not assemble photosystem I. The raw emission spectra obtained from these specimens are quite complex, containing overlapping signatures from many pigments. Multivariate curve resolution analysis of the spectral images shows that each spectrum can be decomposed into independently varying contributions from phycocyanin, allophycocyanin, chlorophyll, a specific pool of "low-energy" chlorophyll associated with photosystem I, and protochlorophyllide. The relative contribution of each of these components varies from species to species in a manner consistent with expectations based on the genetic composition of the mutant strains. For example, we observe significant protochlorophyllide emission from the chlL– mutant which was grown in virtual darkness, while this emission is absent in the wild type. To our knowledge, this is the first ever demonstration of the coupling of rigorous deconvolution methods with hyperspectral confocal microscopy to reveal multiple overlapping native pigment emissions from in vivo specimens. We have also observed evidence for an inhomogeneous distribution of the emitting compounds within the cyanobacterium and are currently quantitatively exploring this inhomogeneity in the concentration distributions both within and between cells. This presentation will describe the design, construction and performance of the hyperspectral confocal microscope. Our results for *Synechocystis* will be described in detail.

Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-ACO4-94AL85000. This work was funded in part by the U.S.Department of Energy's Genomics: GTL program (www.doegenomestolife.org) under project, "Carbon Sequestration in *Synechococcus sp*: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org).

* Presenting author

23

Connecting Temperature and Metabolic Rate to Population Growth Rates in Marine Picophytoplankton

Andrea Belgrano* (ab@ncgr.org) and Damian Gessler

National Center for Genome Resources, Santa Fe, NM

Photosynthetic picophytoplankton bacteria such as *Synechococcus* can contribute up to more then 50% of the total water column primary production, thus playing an important role in controlling the net flux of CO_2 between the atmosphere and the ocean, by the sequestration of carbon from the atmosphere via photosynthesis, and to global carbon cycling in marine systems.

We present how the rate of photosynthesis is regulated by changes in CO_2 concentration, irradiance and temperature, including allometric scaling theory and Rubisco activity as the primary catalyst, for the fixation of carbon by picophytoplankton.

To integrate physiological, biogeochemical, and environmental data in a single model, we integrate ¾-power scaling laws, along with irradiance, temperature, and nutrient uptake functions. Body mass scaling provides the theoretical and empirical model for constraints on the supply and use of energetic resources for picophytoplankton. We use a Boltzamann factor approach to capture the temperature-dependence of metabolism, and additionally introduce a Michaelis-Menten approach for nutrient uptake that includes cell quotas in a single growth model for *Synechococcus*.

We present evidence that a rise in oceanic temperature may reduce the relative contribution that picophytoplankton plays in the ocean as a carbon pathway. This highlights the importance of understanding shifts in the size composition of phytoplankton assemblages in relation to oceanic primary production of biomass, cell density, and nutrient status in the northwestern North Atlantic Ocean.

24

Deciphering Response Networks in Microbial Genomes through Data Mining and Computational Modeling

Z. Su[3], P. Dam[3], V. Olman[3], F. Mao[3], H. Wu[3], X. Chen[1], T. Jiang[1], B. Palenik[2], and Ying Xu[3]* (xyn@bmb.uga.edu)

[1]University of California, Riverside, CA; [2]Scripps Institution of Oceanography, San Diego, CA; and [3]University of Georgia, Athens, GA and Oak Ridge National Laboratory, Oak Ridge, TN

Deciphering of the "wiring diagrams" of biological networks (including metabolic, signaling and regulatory networks) represents a highly challenging problem, due to our lack of general understanding about the conceptual framework of how biomolecules work together as a system and an insufficient amount of experimental data. The majority of on-going computational research has been focusing on developing general methodologies for deriving "functionally equivalent" networks that are consistent with the limited experimental data such as microarray kinetics data, possibly leading to network topologies that are not biologically meaningful.

* Presenting author

We have been developing a computational framework, attempting to systematically derive network topologies that are most consistent with (a) information derived through mining genomic sequences and various genomic and proteomic data and (b) the kinetics data derived from microarray gene expression experiments. The framework consists of the following three key components:

1. **Identification of genes involved in a particular biological process:** To facilitate identification of genes possibly involved in a particular biological network, we first made genome-scale predictions of (a) gene functions, (b) operon structures and (c) *cis* regulatory elements at the genome scale. Gene function prediction is based on available genome annotation plus our own function prediction pipeline using additional information, including motif search and structure-based function prediction. Operons are predicted using our own program (see Section on operon/regulon predictions) *Cis* regulatory elements are predicted using our prediction program CUBIC, in conjunction with microarray data when available, through identification of conserved sequence motifs and similar gene expression patterns. Based on the initial identification of genes possibly involved in a particular biological network, we then refine/expend the gene candidate list through comparing to the information collected in (a), (b) and (c) described above.
2. **Prediction of interaction relationships among these candidate genes:** Currently we attempt to predict two types of interactions: (a) protein-protein interactions, both physical interactions and functional associations, and (b) protein-DNA interactions. Protein-protein interactions are predicted using homology search against protein interaction databases such as DIP & BIND and also based on prediction methods such as gene fusion/fission analysis, and phylogenetic profile analysis. We have developed our own methods for protein (transcription factors)-DNA interactions, based on both sequence and structural information. The sequence-based method is mainly based on (a) homology search against known protein-DNA complexes, (b) identification of self-regulation events, and (c) co-evolution information of transcription factors and operons they regulate. When the 3D structures of transcription factors are available, our method can accurately predict the binding affinity between the structure and its predicted DNA binding motifs, providing a highly effective tool for protein-DNA interaction prediction. Another piece of information for bio-molecular interactions comes from mapping known pathways from related organisms to the target genomes. Though not all such pathway mappings will provide complete pathway models in the target genome, the molecular interactions in the predicted pathways are useful and could be used for piecing together the "complete" network model in (3).
3. **Prediction of wiring diagrams through computational optimization:** We have developed two complementary methods for prediction of "complete" wiring diagrams of a target network, based on the predicted gene candidates and their (partial) interaction relationships and additional information. The first method connects the partially connected pieces predicted in (2) through mapping them to a genome-scale protein-protein interaction map we predicted in (2). The idea is to find the biologically most meaningful "paths" to connect the unconnected pieces (made of protein-protein and protein-DNA interactions). An algorithm has been developed for accomplishing this. In addition, we are currently developing a new algorithm that connects all the interaction components that are most consistent with the available microarray kinetics data, generalizing the current popular methods. By doing so, we can get wiring diagrams that are consistent with both molecular interaction information derived through data mining and microarray kinetics data.

We now describe two key procedures needed to implement the above computational prediction protocol because of the significance by their right.

Prediction of operons and regulons: We have recently developed a computational capability for prediction of operons in microbes, using multiple sources of information including (a) conserved gene neighborhoods across closely related organisms, (b) detected co-evolutionary information of genes, (c) functional relatedness of genes, (d) inter-genic distance information plus various types of other information. The overall prediction accuracy has reached 80% based on our test results on known operons in *E. coli.* We have applied this prediction program, in conjunction with available microarray data, to a number of genomes including *E. coli, Shewanella, Pyrococcus* and *Synechococcus*. The prediction procedure can be outlined as follows. We run the program to produce the initial operon candidate list and then we compare the predicted operons with available microarray data to check for consistency. Corrections will be made on the initial predictions if genes of the same operon exhibit significantly different expression patterns under any experimental condition, or genes from the neighboring operons have highly similar expression patterns under all known conditions and these operons are *very* close in the genomic sequence. In general about 5-10% of the original predictions are corrected based on the microarray data. We expect that the prediction accuracy could reach close to or even beyond 90% when sufficient microarray data is available. Based on the predicted operon structures, we then predicted regulons, based on available microarray data and genome-scale prediction of cis regulatory elements. The prediction procedure identifies operons that share similar expression patterns under the given experimental conditions and share conserved (predicted) binding motifs, and then clusters them into regulons. While this work is still in its early stage, we have identified a number of interesting regulons in the genomes we have applied prediction programs.

Pathway mapping: We have recently developed a computational method and software P-MAP for mapping a known pathway/network from one microbial organism to another by combining homology information and genomic structure information. The basic idea is that in microbes, genes working in the same pathway can generally be decomposed into a few operons or, in case of complex pathways/networks, regulons. Such information has not been effectively used in pathway mapping. When mapping known pathways, we first predict all the operons in a genome using our operon prediction program. The predictions are then validated through comparing microarray data mainly to check for consistency between gene expression patterns for genes predicted to be in the same operons or adjacent operons. Our evaluation has indicated that our prediction accuracy is close to 90%. With such information, we then map genes in a pathway template to the target genome that simultaneously gives relatively high sequence similarity between predicted orthologous gene pairs and has all the mapped genes grouped into a number of operons, preferably co-regulated operons based on the predicted *cis* regulatory elements and available microarray data. We have applied the P-MAP program to map known biological pathways in KEGG and MetaCyc to the cyanobacterial genomes and currently are mapping them to the *Shewanella oneidensis* MR-1 genome. Some of the mapping results could be found at http://csbl.bmg.uga.edu/WH8102.

Applications: We have applied this computational framework to predict the wiring diagrams of various response networks, which consists of signaling, regulatory and metabolic components. These include the carbon fixation, phosphorus assimilation and nitrogen assimilation networks in cyanobacterial genomes. Research is on going to apply the framework to *Shewanella oneidensis* MR-1.

Acknowledgement: This project is supported by the U.S.Department of Energy's Genomics:GTL Program under project "Carbon Sequestration in *Synechococcus sp*: From Molecular Machines to Hierarchical Modeling" (http://www.genomes-to-life.org)

* Presenting author

25

BiLab – A New Tool that Combines the Ease-of-Use of MatLab and the Power of Multiple Computational Biology Libraries

Al Geist* (gst@ornl.gov) and David Jung

Oak Ridge National Laboratory, Oak Ridge, TN

As part of the Genomics:GTL *Synechococcus*: From Molecular Machines to Hierarchical Modeling project, we are developing a new tool called BiLab that we hope will revolutionize computational biology the way the MatLab revolutionized numerical linear algebra. MatLab is widely used to do analysis, to develop new algorithms, and to teach students. MatLab is easy enough for new users and powerful enough for sophisticated users. Yet under the covers it is just a scripting language that provides easy access to the robust linear algebra functions in the LAPACK library. BiLab takes a similar approach except instead of only understanding matrices and doing linear algebra, BiLab understands biological objects such as DNA, proteins, and molecules and is able to manipulate them through any of the functions in a half-dozen standard computational biology libraries. And BiLab is able to display results in biologically relevant form, for example, a protein may be displayed as a molecule, a sequence alignment as stacked sequences.

We developed the BiLab scripting language to cater to different level of expertise in the user, from biologists who just want a quick way to use existing functions to bioinformatics programmers who want to write sophisticated programs in the BiLab scripting language. We have developed the tool to allow the easy addition of new biological objects and functions. Today BiLab provides access to all the functions in bioJava, bioPython, the text based NCBI tools, Jmol, JalView, and CDK. Developers can extend the scripting language to understand new biology. Thus the tool is designed to evolve with the Genomics: GTL program.

Like MatLab, data can be typed in manually or read in from files. BiLab understands the concept of remote biological databases and is able to dynamically load data from SwissProt, GenBank, FASTA, Protein Data Bank, EMBL, and other databases for analysis and study.

This presentation will describe how BiLab is built, how it can be extended, and hands-on demonstrations of the capabilities of the BiLab prototype.

Acknowledgement: This project is supported by the U.S.Department of Energy's Genomics:GTL Program under project "Carbon Sequestration in *Synechococcus sp*: From Molecular Machines to Hierarchical Modeling" (http://www.genomes-to-life.org)

* Presenting author

26

Microbial Cell Modeling via Reacting/Diffusing Particles

Steve Plimpton* (sjplimp@sandia.gov) and Alex Slepoy

Sandia National Laboratories, Albuquerque, NM

We have developed a simulator called ChemCell [1] that tracks protein interactions within cells and can be used to model signaling, metabolic, or regulatory response. Cell features for microbial cells are represented realistically by triangulated membrane surfaces. Particles represent proteins, complexes, or other biomolecules of interest. They diffuse via 3d Brownian motion within the cytoplasm, or in 2d within membrane surfaces. When particles are near each other, they interact in accord with Monte Carlo rules to perform biochemical reactions, which can represent protein complex formation and dissociation events, ligand binding, etc. ChemCell is similar in spirit to MCell [2] and Smoldyn [3].

In this poster, we focus on the underlying algorithms used for reaction rules. We have recently developed a spatial version of the stochastic simulation algorithm (SSA) due to Gillespie [4] and discuss it's implementation in ChemCell. We compare it to alternative approaches including the original SSA and the interaction rules recently developed by Andrews and Bray [3]. We also highlight issues with various reaction/diffusion algorithms relevant to parallel implementation within ChemCell, with the eventual goal of enabling whole-cell models of realistic numbers of proteins and other biomolecules.

References:

1. S. J. Plimpton and A. Slepoy, SAND Report 2003-4509 (2003).
2. J. R. Stiles and T. M. Bartol, in Computational Neuroscience: Realistic modeling for experimentalists, edited by E. De Schutter, published by CRC Press, 87-127 (2001).
3. S. S. Andrews and D. Bray, Phys Biology, 1, 137-151 (2004).
4. D. T. Gillespie, J Comp Phys, 22, 403-434 (1976).

27

Modeling RuBisCO's Gating Mechanism Using Targeted Molecular Dynamics

Paul S. Crozier[1] (pscrozi@sandia.gov), Steven J. Plimpton[1], Mark D. Rintoul[1]*, Christian Burisch[2], and Jürgen Schlitter[2]

[1]Sandia National Laboratories, Albuquerque, NM and [2]Ruhr-Universität Bochum, Bochum, Germany

RuBisCO is the enzymatic bottleneck of carbon sequestration in *Synechococcus*, which is partly due to its catalysis of a competing oxygenase reaction that limits its specificity and efficiency. The binding niche residues are highly conserved across RuBisCO species, yet experimentally-measured specificities and carboxylation rates vary widely. The residues that make up the gate to the binding niche affect the gate's opening and closing rates, and in turn, RuBisCO specificities and reaction rates. We

* Presenting author

have performed molecular dynamics (MD) simulations of RuBisCO's gating mechanism to gain insight into how residue-level changes in RuBisCO's primary sequence affect enzyme performance.

Traditional MD is currently limited to the sub-microsecond timescale, but hardware and algorithm improvements continue to push the attainable timescales upward. We have recently developed several upgrades for our open-source parallel MD simulation package, LAMMPS (http://www.cs.sandia.gov/~sjplimp/lammps.html), which essentially double the algorithm's performance on typical biomolecular simulations. Performance enhancements have come through implementation of the rRESPA hierarchical time-stepping method and a high-performance tabulation algorithm for rapid evaluation of CPU-intensive coulombic interatomic forces. The LAMMPS simulation package was officially released as an open-source parallel MD code available for download on the first of September. Since then, it has been downloaded 1,075 times.

In addition to improving the traditional MD capabilities in LAMMPS, we have implemented advanced MD methods that allow simulation of events that occur on much longer timescales. One such method, targeted molecular dynamics (TMD), allows simulation of user-specified transition events, like the RuBisCO gating mechanism, by imposing a dynamic holonomic constraint on the macromolecular complex. TMD yields the free energy profile of the transition event, which is related to the rate of the transition event.

We performed TMD simulations of the gating event of spinach and *Synechococcus* RuBisCO, each for WT and for mutant D473A. Our simple implicit solvent reduced-model predictions of gating free energy profiles have been encouraging since they have demonstrated the ability to discriminate between RuBisCO structural differences, and are in qualitative agreement with expected trends. For example, our TMD prediction shows a much higher gate opening barrier for *Synechococcus* than for spinach, which indicates more time in the closed state, more photorespiration, and lower specificity for *Synechococcus* RuBisCO. This is in qualitative agreement with the experimentally-measured specificities of *Synechococcus* RuBisCO (47) and spinach RuBisCO (92). Likewise, D473A mutations performed *in silico* for both RuBisCO species show a much lower free energy barrier for gate opening than do wild type RuBisCOs. Experiments show that D473A mutants are not catalytically competent, which is probably due to the fact that the binding niche gate can not properly close (and rapidly opens), without the D473 – R134 salt bridge.

This project is supported by the U.S.Department of Energy's Genomics:GTL Program under project "Carbon Sequestration in *Synechococcus sp*: From Molecular Machines to Hierarchical Modeling" (http://www.genomes2life.org/). Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

28

Selection of Ligands by Panning of Phage Display Peptide Libraries Reveals Potential Partners for TPR Domain and rbcS in *Synechococcus* WH8102

Zhaoduo Zhang* (zzhang@sandia.gov), Arlene D. Gonzales, Todd W. Lane, and Anthony Martino

Sandia National Laboratories, Livermore, CA

One of the goals of functional genomics is the identification of reliable protein interaction partners. The oceanic cyanobacterium *Synechococcus* WH8102 is an abundant marine microorganism important to global CO_2 fixation. We have cloned, expressed and purified two TPR domains of a conserved hypothetical protein and the RuBiscO small subunit protein rbcS from *Synechococcus* WH8102. After immobilizing TPR domains and rbcS, selection of ligands were carried out by panning of two phage libraries displayed random peptides. Peptides specifically binding to TPR domain or rbcS were selected and enriched after three panning processes from a 7-mer and a 12-mer library. A sequence of three amino acids TPR or TPS forms a consensus peptide specific for TPR domains, and APL or APR forms a consensus specific for rbcS. The binding of clones to the target protein was further confirmed by ELISA assay. Peptides specifically binding to rbcS were found in carboxysome protein ccmK2, orfA, csoS3, csoS2 and rbcL, potential partners for rbcS.

University of Massachusetts, Amherst

29

Progress Toward Genome-Scale Monitoring of *In Situ* Gene Expression During Uranium Bioremediation and Electricity Harvesting

Dawn Holmes* (dholmes@microbio.umass.edu), Kelly Nevin, Regina O' Neil, Zhenya Shelbolina, Martin Lanthier, Jonathan Kaye, Brad Postier, and Derek Lovley

University of Massachusetts, Amherst, MA

The first two goals of our Genomics:GTL project are: 1) to determine the genome sequences of the *Geobacteraceae* that predominate during *in situ* bioremediation of uranium-contaminated waters and on the surface of electrodes harvesting electricity from waste organic matter and 2) to comprehensively determine genome-wide patterns of the expression of these *Geobacteraceae* genes in the environments of interest. This data will be used to modify *in silico* models of *Geobacteraceae*, that are initially being developed with data from well-studied pure cultures of *Geobacteraceae* so that these models will be able to more accurately predict the growth and metabolism under a variety of conditions in subsurface environments or on electrodes.

* Presenting author

Substantial progress was made in 2004 in field-scale genomic studies of *in situ* uranium bioremediation and harvesting electricity from aquatic sediments. The *in situ* uranium bioremediation studies were conducted at the DOE-NABIR UMTRA field study site in Rifle, Colorado. Acetate was added to the groundwater in order to stimulate the activity of dissimilatory metal-reducing microorganisms. As previously observed, uranium was rapidly removed from the groundwater as soluble U(VI) was reduced to insoluble U(IV). This is important because it demonstrates the ability to conduct reproducible field experiments at the site.

The composition of the microbial community and groundwater chemistry were monitored daily during the 27 day experiment. The expression of a suite of key genes was monitored every other day. Furthermore, about every 5 days large quantities of groundwater (> 500 liters) were collected for analysis of the genome sequences of the predominant microorganisms. This level of molecular analysis of a subsurface environment is unprecedented.

Analysis of 16S rRNA gene sequences demonstrated that within 13 days *Geobacteraceae* increased from less than 5% of the microbial community when acetate was first injected to over 99% of the community. Quantitative analysis of multiple highly conserved *Geobacteraceae* genes indicated that the number of *Geobacteraceae* increased more than 4 orders of magnitude in this short time. Furthermore, the diversity of *Geobacteraceae* was extremely low. For example, during the height of uranium removal, 82% of the *Geobacteraceae* had 16S rRNA gene sequences that were 97.5-100% identical with 35% having an identical sequence within sequencing error. This is an incredible enrichment of closely related microorganisms in a natural environment. The finding that field experiments can be reproducibly conducted at the Rifle site and that the environment is highly enriched in a small cluster of highly related organisms demonstrates that this environment is ideal for the genome-enabled *in silico* environmental modeling we have proposed.

Furthermore, as the result of our detailed investigations of pure culture *Geobacteraceae* genomes in the first two years of our Genomics:GTL project, it was possible to use newly discovered *Geobacteraceae*-specific gene sequences to conduct detailed studies on levels of *in situ* gene expression in the subsurface. For example, from the analysis of the six available *Geobacteraceae* genomes it was possible to identify *Geobacteraceae*-specific sequences for unique genes such as OmpB, a unique *Geobacteraceae*-specific outer-membrane protein and GltA, a eukaryotic-like citrate synthase unique to *Geobacteraceae*, as well as for genes involved in nutrient uptake and a diversity of housekeeping genes.

Studies on expression of *gltA* in chemostat cultures demonstrated a positive correlation between rates of acetate metabolism and levels of *gltA* transcripts, suggesting that levels of *gltA* transcripts might provide an indication of rates of metabolism. In the field experiment there was a remarkable correspondence between acetate levels in the groundwater and levels of *gltA* transcripts. As acetate rose *gltA* transcript levels increased. Both acetate and *gltA* transcript levels dropped during a rain event that diluted the acetate with rainwater recharge, and then *gltA* levels increased concurrent with a renewed increase in acetate over time. This contrasted with the constant expression, relative to total RNA, of *Geobacteraceae* housekeeping genes such as *recA*, *rpoD*, and *proC*. The pattern of expression of *ompB* was similar to that of the housekeeping genes, consistent with pure culture results indicating that this gene is constitutively expressed and its transcript levels are not correlated with rates of metabolism. Expression of *Geobacteraceae nifD*, which encodes for a portion of the nitrogenase complex, followed a pattern similar to that of *gltA*. This provided further evidence that the metabolism of the *Geobacteraceae* was controlled by the availability of acetate and also demonstrated that the growth of *Geobacteraceae* during bioremediation was limited by the availability of fixed nitrogen.

These results demonstrate that high quality mRNA can be recovered from the subsurface and used to monitor the activity and metabolic state of *Geobacteraceae* during in *situ* uranium bioremediation. In order to monitor gene expression on a genome-wide basis a microarray approach is required. In order to develop this technique, pure cultures of *Geobacter metallireducens* were inoculated into sterilized sediments from the Rifle site which were amended with acetate in order to simulate growth of *Geobacter* species during *in situ* uranium bioremediation. High quality mRNA in sufficient quantities for microarray analysis could be extracted from the sediments and are currently being analyzed with a whole-genome microarray.

In order to construct microarrays that represent the genomes of the *Geobacteraceae* that predominate in the subsurface the genome sequences of these organisms are being determined with three approaches. Isolates with 16S rRNA gene sequences that are identical to those that predominate during *in situ* uranium bioremediation have been recovered and their genomes are being sequenced. High quality genomic DNA was extracted from the subsurface during the 2004 field experiment and is being used to construct both BAC and small insert libraries in order to obtain additional sequence from any *Geobacteraceae* that might not be isolated. Furthermore, samples have been preserved for single-cell genome sequencing. Sequence data from all three approaches should be available at the time of the meeting. This sequence data will be used to construct arrays for genome-scale analysis of gene expression using mRNA extracted for this purpose in the 2004 field experiment.

Field experiments on electricity harvesting were carried out in freshwater and marine sediments at the UMASS field station on Nantucket Island as well as with a swine waste digestor. Electrodes harvesting energy were highly enriched with *Geobacteraceae* of low diversity and representatives of the predominant *Geobacteraceae* were recovered in culture. Studies on gene expression and sequencing genomic DNA similar to those described above for the uranium bioremediation field experiment were carried out and will be presented.

30

Integrating Phenotypic and Expression Data to Characterize Metabolism in *G. sulfurreducens*

R. Mahadevan[1], C. H. Schilling[1], D. Segura[2], B. Yan[3], J. Krushkal[3], and D. R. Lovley[2]*
(dlovley@microbio.umass.edu)

[1]Genomatica, Inc., San Diego, CA; [2]University of Massachusetts, Amherst, MA; and [3]University of Tennessee, Memphis, TN

Geobacteraceae have been shown to be important in bioremediation of uranium contaminated subsurface environments, and in harvesting electricity from waste organic matter. These applications are intricately linked to cellular metabolism, and hence, motivating the need to understand metabolism in these metal reducing bacteria. An iterative approach of mathematical modeling followed by experimentation was adopted to understand metabolism in these organisms.

A genome-scale metabolic model has been developed using the constraint-based modeling approach. Model-based analysis has revealed significant insights on the effect of global proton balance on the physiology of *G. sulfurreducens* and has provided explanation for the reduced yields during Fe (III) reduction. In addition, the comparison of the model predictions of the flux distributions with gene

* Presenting author

expression data was valuable in elucidating the function of genes putatively annotated as encoding for NADPH dehydrogenase. The *in silico* analysis of the energetics of menaquinone secretion indicated a substantial reduction in the growth rate and suggested an explanation for why *Geobacteraceae* predominate over other bacteria that require such electron shuttles. The initial metabolic model provided important physiological and ecological insights on the metabolism of *Geobacteraceae*. However, the analysis of metabolism revealed several redundant pathways in central metabolism around acetate utilization and pyruvate metabolism.

In order to further understand the role of these redundant pathways and their contribution to the overall robustness of metabolism, a combined computational and experimental approach was utilized to unravel the activity of the redundant pathways under different environmental conditions. The computational analysis of the metabolic network identified all the conditionally dependent metabolic pathways. A series of metabolic mutants in pyruvate oxidoreductase (POR), malate dehydrogenase, phosphoenol pyruvate carboxykinase (PPCK), phosphotransacetylase, was designed based on the computational analysis to resolve the activity of the redundant pathways. These mutants were characterized phenotypically under different growth conditions and the experimental data was compared with model predictions. This comparison revealed several interesting aspects of how central metabolism is regulated: POR is the only mechanism for the synthesis of pyruvate from acetate, PPCK is essential for growth on Fe(III) suggesting a potential for this enzyme to be regulated during Fe(III) reduction. These studies indicated the importance of incorporating the mechanism corresponding to the regulation of metabolism to refine the conceptual and *in silico* model.

Gene expression data corresponding to several environmental and genetic perturbations in *G. sulfurreducens* represents information that captures the activity of the regulatory network. Hence, gene expression data derived from several experiments were processed and assembled for further analysis. This expression data was filtered and then clustered based on expression similarities to identify co-expressed genes across the different perturbations. This was followed by sequence analysis including the searching the upstream regions of these co-expressed genes and operons for known transcription factor binding sites, and aligning the upstream regions to identify motifs that correspond to novel sites. This analysis revealed several potential regulatory interactions including a mechanism for regulating heat shock response, and motifs for regulation of sulfate metabolism. Further analysis with additional expression data that incorporates metabolic perturbations is expected to derive regulatory constraints for the metabolic model.

These studies reveal the potential of a combined computational and experimental strategy to iteratively characterize metabolism and the associated regulatory network. Such highly refined conceptual and in silico models of cellular metabolism will be important to design and optimize efficient strategies for bioremediation and harvesting energy from organic substrates.

31

Novel Regulatory Systems and Adaption of Some Well-Known Systems Controlling Respiration, Growth, and Chemotaxis of *Geobactor* Species

Maddalena Coppi[1]* (mcoppi@microbio.umass.edu), Byoung-Chan Kim[1], Laurie DiDonato[1], Julia Krushkal[2], Bin Yan[2], Richard Glaven[1], Regina O' Neil[1], Suphan Bakkal[1], Allen Tsang[1], Hoa Tran[1], Abraham Esteve-Nunez[1], Cinthia Nunez[1], Ching Leang[1], Kuk-Jeong Chin[1], Barbara Methe [3], Robert Weis[1], Pablo Pomposiello[1], Kelly Nevin[1], and Derek Lovley[1]

[1]University of Massachusetts, Amherst, MA; [2]University of Tennessee Health Science Center, Memphis, TN; and [3]The Institute for Genomic Research, Rockville, MD

The goal of our Genomics:GTL project is to be able to predicatively model the growth and activity of the *Geobacter* species that predominate during *in situ* bioremediation of uranium and on the surface of energy-harvesting electrodes in order to better understand these processes and to have the ability to predict the likely outcome of optimization strategies. This requires an understanding not only of the physiological capabilities of *Geobacter* species, but also of how the expression of those physiological capabilities is regulated under various environmental conditions. At last year's meeting we reported on elucidation of global transcriptional regulatory systems in *G. sulfurreducens*, such as the RpoS, RpoE, and Fur regulons. In the past year studies have focused on genome-scale computational and microarray analysis of transcriptional regulation as well as more in depth studies on the regulation of expression of genes known to encode for proteins important in extracellular electron transfer to metals and electrodes.

One of the most surprising findings was that some of the *c*-type cytochromes in *G. sulfurreducens* have a regulatory function and play a role in regulating the production of other *c*-type cytochromes via either transcriptional or post-translation regulatory functions. For example, the *omcF* gene is predicted to encode for a small outer-membrane mono-heme *c*-type cytochrome. Thus, its function as a potential electron transfer protein was evaluated. An OmcF-deficient mutant was deficient in its ability to reduce Fe(III) and this was associated with an absence of the outer-membrane *c*-type cytochrome, OmcB, which is known to be required for Fe(III) reduction. Reverse transcriptase PCR and northern blot analysis revealed that the *omcB* was not transcribed in the OmcF-deficient mutant. Expression of *omcF in trans* restored the expression of *omcB* as well as the ability of the mutant to reduce Fe(III). These results suggest that OmcF may play a role in transcriptional regulation of *omcB*. Deletion of another outer-membrane *c*-type cytochrome gene, designated *omcG*, which is predicted to contain 13 hemes, also greatly diminished levels of OmcB. However, unlike the *omcF* mutant, *omcB* transcription was not affected. These results indicate that OmcG is specifically involved in either the modification, stabilization, or maturation of OmcB. These are the first reports of cytochromes that are necessary for electron transfer to metals, but not directly involved in electron transfer process. Their more likely role is to serve as sensors that regulate cytochrome expression.

Expression of *omcB* is also controlled by RpoS and $\mathrm{Rel}_{\mathrm{Gsu}}$, the *G. sulfurreducens* homolog of RelA, which are important in response to growth under nutrient-limited or stressful conditions. Another transcriptional regulator of *omcB* expression appears to be the product of the gene *ofrR,* which is immediately upstream of the operon that includes *omcB*. Levels of *omcB* transcripts increased orders of magnitude in response to a limitation in electron-acceptor availability or as rates of growth on Fe(III) increased. In a similar manner, expression of *omcS*, which encodes for an outer-membrane *c*-type cytochrome that is required for electricity production is regulated via multiple mechanisms.

* Presenting author

Microarray studies have identified two-component systems that subsequent genetic studies have demonstrated control cytochrome production in response to changing environmental conditions. These results demonstrate that extracellular electron transfer is highly regulated in *G. sulfurreducens*. As outlined in a companion abstract, it has recently been determined that the pili of *G. sulfurreducens* function as nanowires that are required for electron transfer to Fe(III) oxides. Genome-scale studies of the regulation of pilin formation suggested that expression of *pilA*, the gene for the structural pilin protein, is regulated in response to electron acceptor availability, as well as redox and nutrient status. For example, levels of *pilA* transcripts were significantly higher in mutants in which one of the two Fnr-like genes was deleted or when cells were grown under electron-acceptor limiting conditions. Deleting the rel_{Gsu} lowered *pilA* transcript levels. Genome analysis suggests that *pilA* expression is also controlled by a two component regulatory system and the sigma factor, RpoN. Additional mechanisms for pilin production will be reported.

It is expected that regulation of cell behavior in the form of chemotaxis plays an important role in the predominance of *Geobacter* species in subsurface environments. Previous studies have demonstrated that chemotaxis to iron is an important aspect of the reduction of Fe(III) oxides by *Geobacter* species. The genome sequence of *G. sulfurreducens*, contains multiple homologs of chemotaxis genes, including cheW (10), cheA (4), cheY (7), cheR (5), cheB (3), cheC (3), cheD (3) and cheV (1). This contrasts with the genome of *E. coli* which only contains a single copy of a subset of these genes. In order to elucidate factors controlling chemotaxis in *Geobacter* species the gene for a methyl-accepting chemotaxis protein (MCP), from *Geobacter metallireducens* was expressed in a strain of *E. coli* (HCB429) that lacks MCPs. This restored chemotaxis-like in the *E. coli* strain grown in soft agar. Evaluating the function of *Geobacter* chemotaxis proteins in *E. coli* shows promise as a versatile high throughput approach.

Computational analyses that integrated whole genome analyses, comparative genomics, and gene expression microarray data have identified thousands of sites potentially related to gene regulation in *Geobacter* species. A comprehensive resource has been developed that provides the predicted operon organization of the genomes and contig assemblies of *Geobacter* species, potential transcriptional regulatory motifs in the upstream regions of every predicted operon, the results of bi-directional similarity comparisons between *E. coli* regulatory proteins and proteins of *G. sulfurreducens*, and the genome locations of predicted transcription regulatory elements, palindromic motifs, and conserved bacterial elements. This is an invaluable resource for ongoing experimental studies of regulation of respiration and more recently initiated studies on regulation of central metabolism. As will be detailed at the meeting, many of the binding sites for transcriptional regulators that were predicted from analysis of multiple whole-genome gene expression studies with microarrays are in agreement with sequence-based predictions and with experimental results.

Details will also be provided on studies of other forms of regulation such as the response to oxidative stress and the coordinated regulation of the expression of central metabolism and respiratory genes under conditions simulating those expected during *in situ* uranium bioremediation as well as continued studies of the global regulatory systems first described at last years meeting.

32

Nanowires, Capacitors, and Other Novel Electron Transfer Mechanisms in *Geobacter* Species Elucidated from Genome-Scale Investigations

Gemma Reguera[1]* (greguera@microbio.umass.edu), Teena Mehta[1], Dawn E. Holmes[1], Abraham Esteve-Núñez[1], Jessica Butler[1], Barbara Methe[2], Kelly Nevin[1], Swades K. Chaudhuri[1], Richard Glaven[1], Tunde Mester[1], Raymond DiDonato[1], Kevin McCarthy[1], Mark T. Tuominen[1], and Derek Lovley[1]

[1]University of Massachusetts, Amherst, MA and [2]The Institute for Genomic Research, Rockville, MD

Molecular ecology studies have demonstrated that *Geobacteraceae* are the predominant microorganisms involved in *in situ* bioremediation of uranium-contaminated groundwater and on the surface of electrodes harvesting electricity from waste organic matter. However, there has been little information on how these organisms transfer electrons outside the cell onto insoluble electron acceptors, such as metals and electrodes, or alternative mechanisms for respiration.

Insoluble Fe(III) oxides are the primary electron acceptor supporting the growth of *Geobacter* species in subsurface environments, including during uranium bioremediation. Previous studies on electron transfer to Fe(III) oxides in metal-reducing microorganisms have primarily focused on outer-membrane *c*-type cytochromes functioning as the terminal electron carriers that transfer electrons onto Fe(III) oxides. However, analysis of the genomes of two *Pelobacter* species, which are also members of the *Geobacteraceae* indicated that these organisms lacked genes for the outer-membrane cytochromes that are prevalent in *Geobacter* species. Yet, *Pelobacter* species can reduce Fe(III) oxide. If it is assumed that the same mechanism for extracellular electron transfer to Fe(III) oxide is conserved within the *Geobacteraceae*, then these results suggested that outer-membrane *c*-type cytochromes are not the Fe(III) oxide reductase.

Comparison of the available *Geobacteraceae* genomes indicated that the genes for pili, are highly conserved and expression studies indicated that PilA, the structural pilin protein, was expressed during growth on Fe(III) oxide, but not soluble, chelated Fe(III). A mutant of *Geobacter sulfurreducens* in which *pilA* was deleted could reduce soluble electron acceptors, including Fe(III) citrate, as well as wild-type but could not reduce Fe(III) oxide. Complementation with a functional *pilA* restored the capacity for Fe(III) oxide reduction. Based on the role of pili in other organisms, it was hypothesized that the pili were required for *G. sulfurreducens* to attach to Fe(III) oxides, but the *pilA* mutant attached to Fe(III) oxide as well as the wild-type, suggesting a novel role for pili in Fe(III) oxide reduction. Conducting-probe atomic force microscopy revealed that the pili were highly conductive. In contrast, non-pilin proteins had no detectable conductivity and in instances in which the non-pilin proteins covered the pili filaments, they insulated the pili from the conductive tip. No conductivity was detected in the pili of *Shewanella oneidensis*, which, unlike *Geobacter* species, does not need to contact Fe(III) oxides in order to reduce them. These results suggest that the pili of *G. sulfurreducens* serve as biological nanowires, transferring electrons from the cell surface to the surface of Fe(III) oxides. Electron transfer via pili suggests possibilities for other unique cell-surface and cell-cell interactions, and for bioengineering of novel conductive materials.

The finding that pili, rather than *c*-type cytochromes, are likely to be responsible for the final electron transfer to Fe(III) oxides leads to the question of why *c*-type cytochromes are so abundant in

* Presenting author

Geobacter species, both in quantity and number of genes in the genomes. Some *c*-type cytochromes, such as the small, periplasmic PpcA, which are highly conserved in *Geobacter* and *Pelobacter* genomes, may be intermediaries in electron transfer to Fe(III). However, for many outer-membrane cytochromes, there is little similarity in gene sequences in even closely related *Geobacter* species. Whole genome gene microarray and proteomic studies revealed much higher expression of multiple cytochrome genes under electron-acceptor limiting conditions. Further investigation suggested that the cytochromes behave as a capacitor capable of accepting electrons from energy-generating electron transfer reactions in the inner membrane and storing these electrons until a suitable electron acceptor is available. This explains how *Geobacter* species are able to thrive in subsurface environments in which insoluble Fe(III) oxides are heterogeneously dispersed because the capacitor cytochromes permit continued electron transfer during the search for new Fe(III) oxides followed by discharge to the Fe(III) oxide once a suitable source is found.

Surprisingly, the *pilA* mutant produced electricity as well as the wild-type, suggesting that electron transfer to electrodes proceeded via different mechanisms than electron transfer to Fe(III) oxides. Global analysis of gene expression in *G. sulfurreducens* with a whole-genome DNA microarray indicated that the outer-member *c*-type cytochrome gene, *omcS*, and its co-transcribed homolog, *omcT*, were the only genes coding for likely electron transfer proteins that were consistently up-regulated during growth on electrodes versus growth with Fe(III) as the electron acceptor. Quantitative PCR demonstrated that mRNA levels for these cytochromes increased as the amount of current harvested with the electrode increased. When *omcS* and *omcT* were deleted, current production decreased to ca. a third of the wild type and the potential of the anode went from –0.5 V in the wild type to only –0.15 in the mutant. Complementation of the *omcS* gene in the mutant restored current production and anode potential to values comparable to wild type. These results suggest that OmcS is likely to be the primary protein mediating electrical contact between the cell and the electrode surface. This finding offers several possibilities for engineering electrode surfaces and/or microorganisms to improve the function of microbial fuel cells.

Fumarate is an electron acceptor in some *Geobacter* species, such as G. *sulfurreducens*, but not in others, such as *G. metallireducens*. Yet the genomes of both organisms contained what appeared to be a heterotrimic type of fumarate reductase, frdCAB, homologous to the fumarate reductase of *Wolinella succinogenes*. Mutation of the putative catalytic subunit in *G. sulfurreducens* resulted in a strain that lacked fumarate reductase activity and was unable to respire fumarate. Furthermore, the mutant strain could not grow with acetate as the electron donor, regardless of electron acceptor, and lacked succinate dehydrogenase activity. Oxidation of acetate coupled to Fe(III) reduction was possible in the mutant strain if exogenous fumarate was provided, as fumarate could be converted to succinate through TCA cycle reactions and excreted. Highly similar genes were present in *Geobacter metallireducens*, which cannot respire fumarate. When a putative dicarboxylic acid transporter from *G. sulfurreducens* was expressed in *G. metallireducens*, growth with fumarate as the sole electron acceptor was possible. These results demonstrate that, unlike previously described organisms, *Geobacter* species use the same enzyme for both fumarate reduction and succinate oxidation in vivo. This also represents the first example of genetic engineering of a *Geobacter* species for novel respiratory abilities.

Significant progress was also made in genome-enabled studies of oxygen respiration and novel outer-membrane proteins that were first reported at last year's meeting and updates will be provided.

33

Continued Progress in the use of Microarray Technology to Predict Gene Regulation and Function in *Geobacter sulfurreducens*

Barbara Methé[1]*(bmethe@tigr.org), Jennifer Webster[1], Kelly Nevin[2], and Derek Lovley[2]

[1]The Institute for Genomic Research, Rockville, MD and [2]University of Massachusetts, Amherst, MA

Geobacter species represent a rare example in environmental microbiology in which microorganisms closely related to those which predominate in the environment and carry out environmental processes of interest can be readily cultivated in the laboratory. Molecular analyses designed to avoid culture bias, have demonstrated that microorganisms in the family *Geobacteraceae* are the dominant dissimilatory metal-reducing microorganisms in subsurface environments in which organic contaminants are being degraded with the reduction of Fe(III) and in aquatic sediments where dissimilatory metal reduction is important. In addition to their importance in global carbon, nutrient and metal cycles interest in *Geobacter* spp. stems from their potential as agents of bioremediation and capacity to create electricity.

Since completion of the *G. sulfurreducens* genome sequence, global gene expression profiling has been undertaken through the application of microarray technology. Experiments for querying whole genome PCR-based arrays currently being pursued include the examination of wild type *G. sulfurreducens* gene expression profiles under relevant physiological conditions and the testing of mutant strains in which a selected gene has been knocked out versus their wild type counterpart. Various data mining techniques including cluster analyses and analysis of variance are being employed to examine results from individual experiments and collectively across multiple experiments. These efforts have provided new insights into *Geobacter* physiology and regulatory networks.

For example, cells grown with chelated Fe(III) as the electron acceptor had higher levels of transcripts for *omcB* (GSU2737), an outer-membrane *c*-type cytochrome that is essential for Fe(III) reduction. Several other *c*-type cytochrome genes also appeared to be up regulated including a putative *c*-type cytochrome (GSU1334) which based on current genome comparative analyses is unique to *Geobacter* lineages. A substantial proportion (30%) of the significantly expressed genes during Fe(III) reduction were genes of unknown function, or hypothetical proteins. These results suggest differences in the physiology of Fe(III) reduction among microorganisms which perform this metabolic process. An unexpected result was significantly higher levels of transcripts for genes which have a role in metal efflux, potentially suggesting the importance of maintaining metal homeostasis during release of soluble metals when reducing Fe(III). This includes at least six transporter genes that are members of the resistance-nodulation-division (RND) superfamily of efflux transporters such as representatives of the transmembrane spanning heavy metal efflux pump *czc*A family (GSU0830, GSU1332) and one gene encoding for the membrane fusion protein of the *czc*B family (GSU0829). In contrast, transcript levels for other members of the *czc*A and *czc*B families in the *G. sulfurreducens* genome (GSU2135, GSU2136, GSU3400) were comparable during growth on Fe(III) and fumarate. This suggests that these *czc* family members are paralogs with different physiological roles and regulation. Common themes appearing across multiple experiments include the importance of transporter expression and the expression of a group of genes related to protein folding which for example are down regulated under conditions such as growth of *Geobacter* as a biofilm and with Fe(III) as an electron acceptor, in contrast they are up regulated in a mutant strain in which the *rpo*E sigma factor has been knocked out.

* Presenting author

The next phase of the microarray component is building upon these previous successes while extending the flexibility and power of this technique. One example is the adaptation of methods to effect linear amplification of total RNA. Development of this protocol is important to producing high quality hybridizations from samples where the quantity of RNA that can ultimately be obtained from that sample is limited. High quality microarray hybridizations typically require several micrograms of total RNA per replicate. In order to obtain sufficient statistical power for meaningful analyses of microarray data it is necessary to replicate both biological samples as well as within sample replication (technical replication). Cell growth under conditions such as on poorly crystalline iron oxide media (an environmentally relevant growth condition of *G. sulfurreducens*) produces less total RNA upon extraction which can potentially hamper microarray efforts. Linear amplification (as opposed to geometric amplification with traditional PCR) of the total RNA allows the production of sufficient RNA quantities representative of the original proportions of the mRNA transcript population to facilitate the necessary replication of hybridizations to ensure a meaningful outcome post data analysis. The protocol described briefly here is currently being successfully tested on the *G. sulfurreducens* microarray and is a modification of the work of the classic "Eberwine" T7-Amplification method.

Amplified sense RNA is produced using random hexamers in a standard manner for first strand cDNA synthesis to create antisense cDNA. This product (antisense cDNA) is now used as the template in a second strand synthesis along with random nonamers to which a viral T3 promoter is attached. This results in a double-stranded cDNA in which the T3 promoter has been incorporated into the second strand. This second strand is in the sense orientation. An in vitro transcription (IVT) reaction can then be used to transcribe copious quantities of sense RNA from the T3 promoter sites. The resulting IVT product now serves as the template for standard cDNA synthesis and indirect fluorescent (Cy-Dye) labeling. These resulting targets can be used for hybridization to both PCR and oligonucleotide-based arrays.

One example where linear amplification has been applied is to the examination of gene expression patterns of *G. sulfurreducens* when grown using (insoluble) iron oxide as a sole electron acceptor. An overall improvement in hybridization intensity was realized with targets prepared from linear amplified RNA in comparison to unamplified RNA targets. Further, quantitative RT-PCR of a subset of genes from each RNA population (amplified and unamplified) have revealed that the trend of each gene (either significant elevation, depression or no change in gene expression) was the same in both RNA populations. The use of this linear RNA amplification technique in future will include the examination of global gene expression patterns from RNA extracted directly from environments in which *Geobacter* spp. are dominant members of the microbial community.

Shewanella Federation

34

The *Shewanella* Federation: Functional Genomic Investigations of Dissimilatory Metal-Reducing *Shewanella*

James K. Fredrickson*[1](jim.fredrickson@pnl.gov), Carol S. Giometti[2], Eugene Kolker[3]*, Kenneth H. Nealson[4], James M. Tiedje[5], Jizhong Zhou[6], Monica Riley[7], Shimon Weiss[8], James J. Collins[9], Frank Larimer[6], Frank Collart[2], Lee Ann McCue[10], Chip Lawrence[10], and Timothy S. Gardner[9]

[1]Pacific Northwest National Laboratory, Richland, WA; [2]Argonne National Laboratory, Argonne, IL; [3]BIATECH, Bothell, WA; [4]University of Southern California, Los Angeles, CA; [5]Michigan State University, East Lansing, MI; [6]Oak Ridge National Laboratory, Oak Ridge, TN; [7]Marine Biological Laboratory, Woods Hole, MA; [8]University of California, Los Angeles, and [9]Boston University, Boston, MA; and [10]Wadsworth Center, Troy, NY

Shewanella oneidensis MR-1 is a motile, facultative γ-*Proteobacterium* with remarkable metabolic versatility in regards to electron acceptor utilization; it can utilize O_2, nitrate, fumarate, TMAO, DMSO, Mn, Fe, and S^0 as terminal electron acceptors during anaerobic respiration. The ability to effectively reduce nitrate, polyvalent metals including solid phase Fe and Mn oxides and radionuclides such as uranium and technetium has generated considerable interest in the potential role of this organism in metal biogeochemical cycling and bioremediation. The *Shewanella* Federation (SF), a collaborative scientific team assembled by DOE, is applying these approaches to achieve a system-level understanding of how *Shewanella* regulates energy and material flow and to utilize its versatile electron transport system to reduce metals and nitrate. The SF has developed an integrated approach to *Shewanella* functional genomics that capitalizes on the relative strengths, capabilities, and expertise of each group. SF members share information and resources and collaborate on projects that range from a few investigators focused on a defined topic to more complex “Federation-level” efforts. These intend to utilize combined SF capabilities in addressing more general scientific questions.

The SF is organized into various working groups that are engaged in a series of *Shewanella*-based collaborative sub-projects. The ongoing investigations include: **(i)** generation of deletion mutants in each of the 40 predicted *c*-type cytochromes in the MR-1 genome and characterization of their phenotype; **(ii)** global expression profiling using whole genome microarray and proteomic analyses and network modeling of steady-state and transitions between growth on various electron acceptors; and **(iii)** analyses of global regulatory mutants involved in carbon and energy metabolism; and development of common standards and statistical models for numerous SF-wide studies. In addition, coordinate work is underway to revise the annotation of MR-1 based on new information generated by the SF and available in public databases. Supported by the database development, this re-annotation effort will provide the foundation for further functional experimentation, analysis, and modeling of MR-1. Additionally, it will facilitate comparative genomic of analyses of seven new strains of *Shewanella* recently completed by DOE’s Joint Genome Institute. These include *S. amazonensis*, *S. putrefaciens* strains CN32 and 200, *S. denitrificans* OS217T, *S. baltica* OS155, *S. frigidimarina* 400, *S. sp.* str. PV-4. In parallel, detailed physiological analyses are also being conducted with the sequenced strains to provide a basis for genomic comparisons and predictions. The overall goal of this collaborative effort is to develop an evolutionary model for speciation in *Shewanella* as well as providing additional insights into electron transport and carbon metabolism.

* Presenting author

35

Global Profiling of *Shewanella oneidensis* MR-1: Expression of 'Hypothetical' Genes and Improved Functional Annotations

Eugene Kolker*[1] (ekolker@biatech.org), Alex F. Picone[1], Michael Y. Galperin[2], Margaret F. Romine[3], Roger Higdon[1], Kira S. Makarova[2], Natali Kolker[1], Gordon A. Anderson[3], Xiaoyun Qiu[4], Kenneth J. Auberry[3], Gyorgy Babnigg[5], Alex S. Beliaev[3], Paul Edlefsen[1], Dwayne A. Elias[3], Yuri Gorby[3], Ted Holzman[1], Joel Klappenbach[4], Konstantinos T. Konstantinidis[4], Miriam L. Land[6], Mary S. Lipton[3], Lee-Ann McCue[7], Matthew Monroe[3], Ljiljana Pasa-Tolic[3], Grigoriy Pinchuk[3], Samuel Purvine[1,3], Margaret Serres[8], Sasha Tsapin[9], Brian A. Zakrajsek[3], Wenhong Zhu[10], Jizhong Zhou[6], Frank W. Larimer[6], Charles Lawrence[7], Monica Riley[8], Frank R. Collart[5], John R. Yates, III[10], Richard D. Smith[3], Carol Giometti[5], Kenneth Nealson[9], James K. Fredrickson[3], and James M. Tiedje[4]

[1]BIATECH, Bothell, WA; [2]National Institutes of Health, Bethesda, MD; [3]Pacific Northwest National Laboratory, Richland, WA; [4]Michigan State University, East Lansing, MI; [5]Argonne National Laboratory, Argonne, IL; [6]Oak Ridge National Laboratory, Oak Ridge, TN; [7]Wadsworth Center, Albany, NY; [8]Marine Biological Laboratory, Woods Hole, MA; [9]University of Southern California, Los Angeles, CA; and [10]Scripps Research Institute, La Jolla, CA

The γ-proteobacterium *Shewanella oneidensis* strain MR-1 is a metabolically versatile organism that can reduce a wide range of organic compounds, metal ions, and radionuclides. Similar to most other sequenced organisms, approximately 40% of the predicted ORFs in the *S. oneidensis* genome were annotated as uncharacterized 'hypothetical' genes. We implemented an integrative approach using experimental and computational analyses to provide more detailed insight into gene function. Global expression profiles were determined for cells following UV irradiation and under aerobic and suboxic growth conditions. Transcriptomic and proteomic analyses confidently identified 538 'hypothetical' genes as expressed in *S. oneidensis* cells both as mRNAs and proteins (33% of all predicted 'hypothetical' proteins). Publicly available analysis tools and databases and the expression data were applied to improve the annotation of these genes. The annotation results were scored using a seven-category schema that ranked both confidence and precision of the functional assignment. We were able to identify homologs for nearly all of these 'hypothetical' proteins (97%), but could confidently assign exact biochemical functions for only 16 proteins (category 1; 3%). Altogether, computational and experimental evidence provided functional assignments or insights for 240 more genes (categories 2-5; 45%). These functional annotations advance our understanding of genes involved in vital cellular processes including energy conversion, ion transport, secondary metabolism, and signal transduction. We propose that this integrative approach offers a valuable means to undertake the enormous challenge of characterizing the rapidly growing number of 'hypothetical' proteins with each newly sequenced genome.

This work was carried out by the *Shewanella* Federation supported by the *Genomics: GTL* program of the Department of Energy's Offices of Biological and Environmental Research and Advanced Scientific Computing Research (Grants DE-FG02-03ER63527, DE-FG02-02ER63427, and DE-FG02-02ER63342). Argonne, Oak Ridge, and Pacific Northwest National Laboratories are managed for the Department of Energy by University of Chicago (Contract W-31-109-ENG-38), University of Tennessee-Battelle LLC (Contract DE-AC05-00OR22725), and Battelle Memorial Institute (Contract DE-ACO6-76RLO1830), respectively.

* Presenting author

36

Respiratory Pathways and Regulatory Networks of *Shewanella oneidensis* Involved in Energy Metabolism and Environmental Sensing

Alex Beliaev*[1], Yuri Gorby[1], Margie Romine[1], Jeff McLean[1], Grigoriy Pinchuk[1], Eric Hill[1], Jim Fredrickson[1], Jizhong Zhou[2], and Daad A. Saffarini[3]

[1]Pacific Northwest National Laboratory, Richland, WA; [2]Oak Ridge National Laboratory, Oak Ridge, TN; and [3]University of Wisconsin, Milwaukee, WI

Shewanella oneidensis MR-1 is a facultative γ-Proteobacterium with remarkable metabolic versatility in regards to electron acceptor utilization; it can utilize O_2, nitrate, fumarate, Mn, Fe, and S^0 as terminal electron acceptors during respiration. This versatility allows MR-1 to efficiently compete for resources in environments where electron acceptor type and concentration fluctuate in space and time. The ability to effectively reduce polyvalent metals and radionuclides, including solid phase Fe and Mn oxides, has generated considerable interest in the potential role of this organism in biogeochemical cycling and in the bioremediation of contaminant metals and radionuclides. The entire genome sequence of MR-1 has been determined and high throughput methods for measuring gene expression are being developed and applied. This project is part of the *Shewanella* Federation, a multi-investigator and cross-institutional consortium formed to achieve a systems level understanding of how *S. oneidensis* MR-1 senses and responds to its environment.

Electron Acceptor-Induced Shifts in *S. oneidensis* MR-1 Gene Expression Profiles. To define the repertoire of genes responding to both metal and non-metal electron acceptors and identify basic regulatory mechanisms governing anaerobic respiration in *S. oneidensis*, we compared mRNA expression patterns of anaerobic cultures incubated with fumarate to those exposed to nitrate, thiosulfate, DMSO, TMAO, ferric citrate, hydrous HFO, manganese dioxide, colloidal manganese, and cobalt using whole-genome arrays. The extent of *S. oneidensis* transcriptome response to metal electron acceptors was revealed by hierarchical clustering analysis, where a high degree of similarity in global expression profiles was exhibited throughout all metal-reducing conditions which resulted in metals grouping separately from non-metal electron acceptors and forming a tight, well-defined branch. In accordance with the results of a principal component analysis, we identified two major expression groups that displayed activation and repression in the presence of metals and accounted for over 60% of the differentially expressed genes. While genes encoding hypothetical and conserved hypothetical proteins dominated both clusters, there were several functional subgroups encoding putative components of electron transport chain, transcriptional regulators and detoxification/toxin resistance proteins that were characterized by their non-specific upregulation to all metal electron acceptors. Contrary to what was expected, the *mtrCAB* operon which encodes two deca-heme *c*-type cytochromes and an outer membrane protein essential for Fe(III) and Mn(IV) respiration in *S. oneidensis* showed 2- to 8- fold decrease in mRNA levels under metal-reducing conditions. In contrast, *S. oneidensis* demonstrated specific transcriptome responses to individual non-metal electron acceptors producing unique clusters of nitrate, thiosulfate and TMAO induced genes. While these observations undoubtedly reflect the nature of metal and non-metal electron acceptors, the diversification and tighter regulatory control of the non-metal respiratory systems may be indicative of different evolutionary pathways taken by these respiratory systems. Moreover, the absence of upregulation for known genes involved in metal reduction may be due to the low-specificity and the opportunistic nature of the metal reduction pathways in *S. oneidensis*. This work represents an important step to-

* Presenting author

wards understanding the anaerobic respiratory system of *S. oneidensis* MR-1 on a genomic scale and has yielded numerous candidate genes for more detailed functional analysis.

Autoaggregation of *Shewanella oneidensis* in Response to High Oxygen Concentrations. Despite the potential environmental importance of this phenomenon, little is known about the mechanisms inducing aggregate formation and subsequent impacts on cells inside the aggregates. Under aerobic conditions, *S. oneidensis* cells are highly adhesive to glass and in the presence of $CaCl_2$ the cells aggregate into large multi-cellular structures. Microscopic analyses of these aggregates identified a presence of DNA, proteins and carbohydrate-like material in the extracellular matrix. In contrast, cells grown under suboxic conditions did not display any autoaggregation while their adhesion to surfaces was significantly reduced. Microarray expression analysis comparing samples of suboxically- vs. aerobically-grown cells identified a set of genes encoding cell-to-cell and cell-to-surface adhesion and colonization factors that positively responded to increased O_2 concentrations. Of particular interest was the O_2-dependent upregulation of *S. oneidensis csgAB* and *csgDEF* operons which are putatively involved in curli fimbrae formation. In other organisms, such *Escherichia coli*, these structures confer attachment to inert surfaces such as glass and have also been implicated in cell-cell attachment. Although, when compared to suboxic conditions, both flocculated and unflocculated cells displayed some similarities in gene expression in response to elevated levels of O_2, autoaggregation had a significant impact on gene expression in *S. oneidensis*. Direct comparison of aggregated versus unagreggated cells grown under 50% dissolved O_2 tension (DOT) revealed remarkable differences in mRNA patterns between these two states. Unflocculated cells displayed significant increase of mRNA levels of genes involved in aerobic energy metabolism, intermediary carbon metabolism and gluconeogenesis as well as chemotaxis and motility. In contrast, several genes putatively involved in anaerobic metabolism, gene cluster encoding outer membrane proteins and cytochromes, and transcriptional regulation were upregulated under 50% DOT aggregated conditions. Remarkably, the majority (~90%) of genes located on the 50-kb megaplasmid of MR-1 displayed substantial levels of upregulation in flocculated cells. It is currently unclear whether this phenomenon is due to a global regulatory effect or to an increase in plasmid copy number. Although further studies are required for resolution, we speculate that autoaggregation in *S. oneidensis* MR-1 may serve as a mechanism to facilitate reduced O_2 tensions within aggregate, leading to the expression of anaerobic genes under bulk aerobic conditions.

Cyclic AMP Signaling and cAMP Receptor Protein-Dependent Regulation of Anaerobic Energy Metabolism in *Shewanella oneidensis* MR-1. Unlike many bacteria studied to date, the ability of *S. oneidensis* to grow anaerobically with several electron acceptors is regulated by the cAMP-receptor protein (CRP). CRP-deficient mutants of MR-1 are impaired in anaerobic reduction and growth with Fe(III), Mn(IV), fumarate, nitrate, and DMSO. Loss of anaerobic respiration in *crp* mutants is due to loss of terminal anaerobic reductases and not due to deficiency in carbon metabolism. To further elucidate the role of CRP and to understand the mechanisms of cAMP-dependent gene expression under anaerobic conditions in *S. oneidensis*, DNA microarray analyses were performed. Comparison of mRNA expression profiles of wild-type and *crp* mutant cells grown anaerobically with different electron acceptors indicated that CRP positively regulates the expression of genes involved in energy generation and transcriptional regulation. These include the periplasmic nitrate reductase, the polysulfide reductase, anaerobic DMSO reductase genes, as well as the nitrate/nitrite sensor protein *narQ*. To identify the mechanisms and proteins that lead to CRP activation under anaerobic conditions, genes predicted to encode adenylate cyclases were analyzed. The genome sequence of *S. oneidensis* contains three putative adenylate cyclase genes, designated *cyaA*, *cyaB*, and *cyaC*. Deletion of *cyaA* or *cyaB* did not affect anaerobic growth with any of the electron acceptors tested while deletion of *cyaC* resulted in growth deficiency with DMSO. Surprisingly, deletions of both *cyaA* and *cyaC* resulted in anaerobic growth deficiency with DMSO, nitrate, Fe(III), Mn(IV), and fumarate. These phenotypes are similar to the phenotypes of the CRP-deficient mutants. Our results indicate that both CyaA and CyaC are needed for the production of cAMP under anaerobic conditions, and for activation of CRP. Further work to identify the cAMP signaling pathways in *S. oneidensis* is underway.

37

Functional Analysis of *Shewanella*, A Cross Genome Comparison

Margrethe H. Serres* (mserres@mbl.edu) and Monica Riley

Marine Biological Laboratory, Woods Hole, MA

The genome sequence of *Shewanella oneidensis* MR-1, a microbe with unique metabolic and respiratory properties including the use of metals as electron acceptors, has been published (Heidelberg et al. 2002). Recently six additional *Shewanella* genomes; *S. putrefaciens* CN-32, *S. alga* PV-4, *S. amazonensis* SB2B, *S. baltica* OS155, *S. frigidimarina* NCIMB400, and *S. denitrificans* OS2717 have been sequenced at the Joint Genome Institute. The new strains were selected as representatives of different ecological niches and redox environments ranging from terrestrial to marine and freshwater sediments. We are making use of the new sequences to do a comparative analysis between the protein coding sequences of *S. oneidensis* and MR-1 and those of the newly available *Shewanella* strains. In addition we are including the genomes of *Escherichia coli*, an experimentally well studied organism, and *Geobacter sulfurreducens*, an organism with metal reducing capabilities. Our aim is to detect variations in the protein content that may be related to differences in metabolic properties and environmental adaptation for the organisms.

In our previous work we have assembled groups of sequence similar proteins of *S. oneidensis* MR-1. We initially identified fused proteins (155) as these complicate the grouping process and separated them into stand-alone functional entities. The Darwin algorithm was used to detect sequence similarity among the protein sequences. A transitive grouping process was applied to generate sequence similar groups containing both closely related as well as more distantly related members. We also restrict the membership of a protein to one sequence similar group. As a result we identified 406 paralogous or sequence similar groups with memberships ranging from 2 to 64. The largest paralogous groups were found to encode for response regulators, ATP-binding component of the ABC superfamily transporters, transcriptional regulators of the LysR family, regulatory proteins, and sensory histidine kinases. The group sizes show a power-law distribution with most of the groups having few members and few groups having many members. The paralogous groups represent ancestral genes which have gone through duplication and divergence to generate today's gene families. These gene families encode for proteins with related functions. Gene duplication and divergence is thought of as an important means by which an organism may specialize or generate functions.

We are making use of the paralogous group data for the *S. oneidensis* MR-1 genome to search for sequence matches in the newly sequenced *Shewanella* genomes and in *E. coli* and *G. sulfurreducens*. Data will be presented on the distribution of homologs to the members of the *S. oneidensis* MR-1 paralogous groups. The data will be analyzed to identify differences or similarities between the or-

* Presenting author

ganisms and further use this information to shed light on functions that may be of importance to the metabolic properties and environmental fitness of the organisms.

One of the larger *S. oneidensis* MR-1 paralogous groups was found to contain a family of 27 methyl-accepting chemotaxis proteins (MCPs). These proteins are located in the membrane where they serve as signal receptors for the chemotaxis apparatus. Five MCPs are encoded in the *E. coli* genome and they are known to bind specific attractants or repellants. The expanded repertoire of MCPs in *S. oneidensis* MR-1 suggests that the chemotactic response and environmental sensing is highly specialized in this organism. Sequence similarity between the *S. oneidensis* MR-1 MCPs and the proteins encoded in the newly sequenced *Shewanella* strains shows an expansion of the MCPs versus that of *E. coli* for all the strains. Interestingly the number of proteins with sequence matches varies widely in the different genomes from 19 to 40. *G. sulfurreducens* appears to contain 33 homologs of the *S. oneidensis* MR-1 MCP group. The sequence similarity and distribution of other chemotaxis related proteins will be presented.

38

Optical Methods for Characterization of Expression Levels and Protein-Protein Interactions in *Shewanella oneidensis* MR-1

Natalie R. Gassman[1]* (ngassman@chem.ucla.edu), Xiangxu Kong[1], Gopal Iyer[1], Younggyu Kim[1], and Shimon Weiss[1,2]

[1]University of California, Los Angeles, CA and [2]University of California, Santa Barbara, CA

Biological networks are dependent on a delicate balance of cellular signaling and dynamic transcriptional response, and it has become increasingly important to unravel these networks by accurately quantifying gene expression levels and mapping protein-protein interactions. We have developed a single optical technique, Alternating Laser Excitation (ALEX), which can integrate the analysis of protein-protein interactions and gene expression levels in a sensitive and potentially high-throughput manner.

A protein-protein interaction rich system in MR-1 is transcriptional regulation. Using the sigma factor, σ^{24}, and the DNA bending protein IHF, we are currently reconstructing an active transcription system from MR-1 and creating fluorescently labeled proteins for interaction analysis. When fluorescently labeled proteins are characterized, the ALEX method can be used to examine the mechanistic process of gene regulation by σ^{24} -RNA polymerase (RNAP) from the formation of open complex to transcription elongation, and the dynamics of DNA bending by IHF for transcriptional initiation can be resolved.

Additionally, we are currently expanding the capabilities of the ALEX technique for gene expression analysis. To demonstrate advances in this effort, we examined a DNA model system using two-color coincident detection. By hybridizing two spectrally distinct fluorophores to the target, we detected and quantified individual DNA molecules. Hybridized complexes were detected by colocalization of the probes to the target and specificity of the probes was determined by Forster resonance energy transfer (FRET) between the probes. With ALEX, we have detected and quantified interactions at the DNA level and are currently focusing our efforts at the mRNA level. Progress in both quantification of gene expression and protein-protein interactions will be reported.

39

Reverse-Engineering Microbial Networks in *Escherichia coli* and *Shewanella oneidensis* MR-1 via Large-Scale Perturbation Studies

G. Cottarel, M.E. Driscoll, J. Faith, M.K. Kohanski, J. Wierzbowski, C.B. Cantor, J.J. Collins, and T.S. Gardner* (tgardner@bu.edu)

Boston University, Boston, MA

The impressive capabilities of microbes, ranging from energy transduction to signal processing, rival those of any engineered system. Functions of respiration, growth, and environmental sensing are principally regulated by transcriptional gene networks. Identifying the large-scale structure and dynamics of such networks is an important first step towards engineering microbes for applications in bioremediation and energy production.

Towards this goal, we have recently completed a pilot study validating an approach for rapid cell-wide reverse-engineering of transcriptional gene networks. In an extended study of the DNA-damage response network of *Escherichia coli*, we generated genome expression profiles of cells under 65 experimental conditions, encompassing both time-series profiles and genetic perturbations, in a background of antibiotic-induced DNA damage.

We succeeded in reconstructing a network map comprising over one hundred genes using an inference algorithm developed previously in our lab [1,2]. This network provides a comprehensive picture of a major stress-response system in prokaryotes, buttressing and unifying evidence from previous studies. In addition, we have also identified several novel regulators in the network, for which we are pursuing further experimental validation. Our results establish the feasibility and scalability of our reverse-engineering approach, and have laid the groundwork for a similar study in *Shewanella oneidensis* MR-1.

S. oneidensis is a gram-negative microbe whose ability to reduce heavy-metals and other organic toxins has made it a promising candidate for use in environmental remediation. Using our reverse-engineering methods, we are conducting a broad series of growth-condition perturbations to reconstruct the transcriptional networks governing *Shewanella*'s respiratory system.To facilitate these studies, we have designed the first high-density Affymetrix oligonucleotide microarray for *S. oneidensis*. Beyond its use for genome-wide expression profiling, this microarray can also be used in chromatin immunoprecipitation studies, which will complement our expression-based inference techniques.

The knowledge we are deriving from our work both in *E. coli* and *S. oneidensis* has a variety of applications, including the improvement of antibiotics, environmental remediation, and the prospect of biologically-derived energy sources.

References

1. Gardner TS, di Bernardo D, Lorenz D, Collins JJ. Inferring genetic networks and identifying compound mode of action via expression profiling. Science. 2003 Jul 4;301(5629):102-5.
2. Tegner J, Yeung MK, Hasty J, Collins JJ. Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling. Proc Natl Acad Sci U S A. 2003 May 13;100(10):5944-9.

* Presenting author

40

Comparative Analysis of Gene Expression Profiles of *Shewanella oneidensis* MR-1 Following Exposure to Ionizing Radiation and Ultraviolet Radiation

Xiaoyun Qiu[1]* (qiuxiaoy@msu.edu), George Sundin[1], Michael J. Daly[2], Alexander Vasilenko[2], Marina V. Omelchenko[3], Jizhong Zhou[4], Liyou Wu[4], Mary S. Lipton[5], and James M. Tiedje[1]

[1]Michigan State University, East Lansing, MI; [2]Uniformed Services University of the Health Sciences, Bethesda, MD; [3]National Institutes of Health, Bethesda, MD; [4]Oak Ridge National Laboratory, Oak Ridge, TN; and [5]Pacific Northwest National Laboratory, Richland, WA

Shewanella oneidensis MR-1, a Gamma proteobacterium, is notable in the terminal electron acceptors it uses including some toxic metal ions and radionuclides. Thus it has potential for bioremediation. However, MR-1 is highly sensitive to both ionizing radiation (IR) and ultraviolet radiation (UVR). We delineated the genomic response of *Shewanella oneidensis* MR-1 to gamma ray, UVC (254 nm), UVB (290-320 nm), UVA (320-400 nm) and natural solar radiation. A total of 5.9-, 4.2-, 3.9-, 8.1-, and 28.0% of the MR-1 genome showed differential expression ($P<0.05$ and fold >2), respectively, at a dose that yielded about 20% survival rate. The gene expression profile of MR-1 in response to ionizing radiation is more similar to that of UVC, which is characterized by the induction of SOS response and prophage synthesis, plus a strong induction of antioxidant enzymes. Genomic response to UVB is a combination of the UVC and UVA patterns, which represents a shift from shorter wavelength of UVR-induced direct DNA damage and activation of prophages to longer wavelength of UVR-induced global photo-oxidative damage. We observed the traditional UVA-induced stress responses in MR-1 such as induction of antioxidant enzymes and proteins, sequestration of the transition metals and activation of the degradative pathways, however, the induction of heavy metal and multidrug efflux pumps is a previously unknown phenotype for this stress. Consistent with natural solar UV radiation composition (about 95% UVA and 5% UVB), genomic response to solar radiation is more similar to that of UVA but with more genes induced for detoxification. In addition, the number of differentially expressed genes from most functional categories was much greater than for UVB or UVA or their sum. This unique gene expression profile indicates that natural solar radiation impacts biological processes in a much more complex way than previously thought.

Comparative genome analysis indicates that *S. oneidensis* MR-1 encodes a complex set of DNA repair and detoxification genes. For example, about 2.8% of the MR-1 genome is dedicated to DNA repair, replication and recombination, which is very similar to that of *Escherichia coli* K12 (2.7%) and the extremely radiation resistant *Deinococcus radiodurans* R1 (3.1%). However, only about 5.8- and 13.9% of those genes were induced in MR-1 by UVC and gamma ray, respectively, which is much lower than in *E. coli* K12 (15.7% were induced by UVC) and in *D. radiodurans* (22.0% were induced by gamma ray). This result indicates that alteration in gene content and gene regulation, which may be the consequence of lack of recent natural selection, contribute to the high radiation sensitivity of MR-1.

Although we observed a strong induction of the SOS response in MR-1 following exposure to IR or short wavelength UVR (UVC and UVB), DNA damage caused during irradiation itself might not be the primary cause of cell death since there is relatively little DNA damage in MR-1 following 40 Gy or 3.3 J m^{-2} of UVC. MR-1 is a respiratory generalist and is very rich in cytochromes, which together with other respiratory chain components such as flavins and quinones are an important

* Presenting author

source of oxidative stress. A recent study by Daly et al. showed that in contrast to *D. radiodurans*, MR-1 accumulates exceptionally high Fe and low Mn levels. Accumulation of Mn in bacteria has been proposed to serve as non-enzymatic antioxidants during recovery after radiation exposure. Collectively, the results support that irradiated *S. oneidensis* is responding to oxidative stress elicited by metabolism-induced free radicals produced during recovery. For Fe-rich, Mn-poor cells such as *S. oneidensis*, death at low doses of IR might be caused by a combination of oxidative stress and the induction of lytic prophages, as proposed following recovery from UV radiation.

41

The Microbial Proteome Project: A Database of Microbial Protein Expression in the Context of Genome Analysis

Carol S. Giometti [1]* (csgiometti@anl.gov), Gyorgy Babnigg[1], Sandra L. Tollaksen[1], Tripti Khare[1], Angela Ahrendt[1], Wenhong Zhu[2], Derek R. Lovley[3], James K. Fredrickson[4], and John R. Yates III[2]

[1]Argonne National Laboratory, Argonne, IL; [2]Scripps Research Institute, La Jolla, CA; [3]University of Massachusetts, Amherst, MA; and [4]Pacific Northwest National Laboratory, Richland, WA

This project, funded through the Microbial Genome and Genomics:GTL programs, is focused on the detection and characterization of differential protein expression in microbial systems relevant to the goals of the Office of Biological and Environmental Research. As part of this effort, relational databases are being used to assimilate and integrate the data collected. This growing knowledgebase can serve as a resource for comparison of proteomes across species and assessment of differential cellular responses to a variety of growth conditions. The Microbial Proteome Project knowledgebase currently includes experimental details and proteome data for 11 different microbes (*Deinococcus radiodurans, Geobacter sulfurreducens, Geobacter metallireducens, Methanococcus jannaschii, Prochlorococcus marinus, Pyrococcus furiosus, Psychrobacter sp.5*, *Rhodopseudomonas palustris, Rhodobacter sphaeroides, Shewanella oneidensis,* and *Synechocystis sp. PC*) and includes over 8000 protein pattern images that are accessible to authenticated users.

As partners in the *Shewanella* Federation and University of Massachusetts Genomics:GTL projects, our laboratory efforts in the past year have been focused on the proteomes of *S. oneidensis MR-1*, *Geobacter sulfurreducens*, and *Geobacter metallireducens*. Using two-dimensional electrophoresis (2DE) analyses, we have been able to complement the results generated using LC/MS-MS proteomics and microarray mRNA analysis methods, providing "snapshots" of protein components to confirm the expression of specific proteins, measure their relative abundance, and detect post-translational modifications. In addition to traditional 2DE methods, we have introduced approaches that capitalize on the sensitivity of fluorescence and chemiluminescence. Using these approaches, we have been able to detect differences in protein phosphorylation patterns and in heme protein expression. By combining affinity chromatography to enrich samples for phosphoproteins with detection using phosphate-specific fluorescent dyes and immunoblotting with antibodies against specific phosphoamino acids, we have detected differences in the phosphoproteome of *S. oneidensis* when cells are grown aerobically compared to anaerobically. To optimize our detection of *c*-type cytochromes, we have used an isoelectric focusing fractionation protocol to concentrate *G. sulfurreducens* and *G. metallireducens* protein samples prior to electrophoresis and then detected the heme-positive proteins using a chemiluminescence assay. This latter method has increased the number of heme-positive proteins detectable from

* Presenting author

only four or five up to approximately 25 for each *Geobacter* species, and has provided the opportunity to do comparative analysis of the *c*-type cytochromes expressed by each of these microbes grown under different conditions.

In parallel with our investigation of subsets of the microbial proteomes, we continued our efforts to identify both constitutively expressed and induced proteins in *S. oneidensis*, *G. sulfurreducens*, and *G. metallireducens*. Our knowledgebase currently includes the identifications of 575 *S. oneidensis* MR-1, 400 *G. sulfurreducens* PCA, and 77 *G. metallireducens* proteins together with the peptide mass data used to obtain these identifications. These identifications represent the most abundant protein spots detected in 2DE patterns with either Coomassie Blue R250 or silver nitrate and include proteins expressed in *S. oneidensis* cells grown with 50% dissolved oxygen (i.e., aerobic conditions), 0% dissolved oxygen (i.e., suboxic or hypoxic), Fe(III), nitrate, or fumarate, *G. sulfurreducens* cells grown with either fumarate or Fe(III), and *G. metallireducens* cells grown with either nitrate or Fe(III). When the functional categories of the identified *S. oneidensis* and *G. sulfurreducens* proteins are compared, the relative number of identified proteins in each category is quite similar. Exceptions are found in the categories of cellular processes and protein fate with a greater percentage of *S. oneidensis* proteins identified than *G. sulfurreducens* proteins. More proteins associated with regulatory functions, on the other hand, have been identified among the *G. sulfurreducens* proteins detected in 2DE patterns than among the *S. oneidensis* proteins. Also, a larger number of *G. sulfurreducens* proteins with no annotation or annotated as having unknown function have been identified whereas more proteins annotated as hypothetical proteins has been identified in the 2DE patterns of *S. oneidensis*. The latter observation could reflect the less mature annotation of the *G. sulfurreducens* genome compared to the *S. oneidensis* genome, i.e., proteins currently having no functional annotation could be reannotated as proteins with known function as improved methods for functional annotations become available. Of the 77 *G. metallireducens* proteins identified, 25 are currently annotated as unknown, hypothetical, or conserved hypothetical proteins.

Since the overall goal of this project is to provide a public resource of protein expression information for microbes in the context of genome sequence, in addition to a secure website shared by *Shewanella* and *Geobacter* co-investigators, public websites have been designed to provide access to proteome analysis results as they become validated and published. ProteomeWeb (http://ProteomeWeb.anl.gov) is an interactive public site that currently provides the identification of *M. jannaschii* proteins detected by 2DE with links to genome sequence information, tools for mining the proteome data, and links to metabolic pathways. GelBank (http://GelBank.anl.gov) includes the complete genome sequences of approximately 205 microbes and is designed to allow queries of proteome information. Numerous tools are provided, including the capability to search available sequence databases for specific protein functions and amino acid sequences. Web applications pertinent to 2DE analysis are provided on this site (e.g., titration curves for collections of proteins, 2DE pattern animations). The database is currently populated with protein identifications from the Argonne Microbial Proteomics studies and will accept data input from outside users interested in sharing and comparing results from proteome experiments.

This work is supported under Contract No. W-31-109-ENG-38 with the U.S. Department of Energy.

J. Craig Venter Institute

42

Estimation of the Minimal Mycoplasma Gene Set Using Global Transposon Mutagenesis and Comparative Genomics

John I. Glass* (JGlass@venterinstitute.org), Nina Alperovich, Nacyra Assad-Garcia, Shibu Yooseph, Mahir Maruf, Carole Lartigue, Cynthia Pfannkoch, Clyde A. Hutchison III, Hamilton O. Smith, and J. Craig Venter

J. Craig Venter Institute, Rockville, MD

The Venter Institute aspires to make bacteria with specific metabolic capabilities encoded by artificial genomes. To achieve this we must develop technologies and strategies for creating bacterial cells from constituent parts of either biological or synthetic origin. Determining the minimal gene set needed for a functioning bacterial genome in a defined laboratory environment is a necessary step towards our goal. For our initial rationally designed cell we plan to synthesize a genome based on a mycoplasma blueprint (mycoplasma being the common name for the class *Mollicutes*). We chose this bacterial taxon because its members already have small, near minimal genomes that encode limited metabolic capacity and complexity. We took two approaches to determine what genes would need to be included in a truly minimal synthetic chromosome of a planned *Mycoplasma laboratorium*: determination of all the non-essential genes through random transposon mutagenesis of model mycoplasma species, *Mycoplasma genitalium*, and comparative genomics of a set of 15 mycoplasma genomes in order to identify genes common to all members of the taxon.

Global transposon mutagenesis has been used to predict the essential gene set for a number of bacteria. In *Bacillus subtilis* all but 271 of bacterium's ~4100 genes could be knocked out. Because of the redundancy of many genes and systems in *B. subtilis* and other conventional bacteria, in these bacteria disruptions of genes involved in essential functions are often not lethal. *M. genitalium* is a slow-growing human urogenital pathogen that has the smallest known genome of any free-living cell at 580 kb. There is almost no redundancy in this genome, and as such *M. genitalium* is often used as a model of a minimal cell. In 1999 we published preliminary estimate of the *M. genitalium* minimal gene set. Global transposon mutagenesis identified 130 of the 484 *M. genitalium* protein-coding genes not essential for cell growth under laboratory conditions, and that work predicted that in a complete study still additional genes would likely be disrupted. In our current effort we have improved the technique to allow isolation and characterization of disruption mutants. Surprisingly, after attaining saturation mutagenesis of the *M. genitalium* genome we could only identify 98 disrupted genes, suggesting that for growth under our laboratory conditions the minimal mycoplasma essential gene set is ~386 protein coding genes. Some genes were disrupted that are involved in presumably critical metabolic processes, such as lactate, pyruvate and glycerol-3-phosphate dehydrogenases. This suggests that as has already been shown for some *M. genitalium* kinases, these dehydrogenases may be somewhat redundant due to less than stringent substrate specificity.

The 15 mycoplasma genomes comprise an excellent comparative genomics virtual laboratory. Previous similar computational comparisons of genomes across diverse phyla of the eubacteria are of limited value. Because of non-orthologous gene displacement, pan-bacterial comparisons identified

* Presenting author

less than 100 genes common to all bacteria; however determination of conserved genes within the narrow mycoplasma taxon is much more instructive. We identified 169 protein coding genes present in all of the complete mycoplasma genome sequences, and an expanded core set of 310 genes that are encoded in almost every member of our set of genomes. The additional genes in the expanded core gene set take into account that non-glycolytic mycoplasmas do not encode some glycolysis genes for instance, and that the obligate intracellular plant parasite, *Phytoplasma asteris*, has dispensed with many genes because the functions are provided by its host. At least 36 elements of the expanded core gene set are non-essential based on the gene disruption study. The combination of comparative genomics with the gene disruption data, and reports of specific enzymatic activities in different mycoplasma species enabled us to predict what elements are critical for this bacterial taxon. In addition to determining the consensus set of genes involved in different cellular functions, we identified 10 hypothetical genes conserved in almost all the genomes, and paralogous gene families likely involved in antigenic variation that comprise significant fractions of each genome and are presumably unnecessary for cell viability under laboratory conditions.

43

Progress toward a Synthetic Cellular Genome

Hamilton O. Smith* (hsmith@venterinstitute.org), Cynthia Pfannkoch, Holly A. Baden-Tillson, Clyde A. Hutchison III, and J. Craig Venter

J. Craig Venter Institute, Rockville, MD

To test our understanding of the genetic requirements for cellular life, we proposed to construct a minimal cellular genome by chemical synthesis (Hutchison, et al., 1999). A number of technical hurdles remain before this can be accomplished and we report progress on these here.

We improved upon the methodology and shortened the time required for accurate assembly of 5- to 6-kb segments of DNA from synthetic oligonucleotides. We first tested our methodology by assembly of infectious phiX174 genomes (Smith, et al., 2003). The methods have since been tested by assembling three segments (4.6, 5.3, and 6.5 kb) of the mouse mitochondrial genome. In each case assembly was straightforward, so we feel the methods are quite robust. We have made improvements in these assembly methods aimed at improving sequence accuracy of the assembled DNA.

About one lethal error per 500 bp occurred in our phiX174 genome assembly, assuming a random distribution of errors among assembled genomes. We have approached the elimination of such errors at four levels: 1) Purifying oligonucleotides prior to assembly by high-throughput capillary electrophoresis, to remove molecules containing errors that lead to mutations 2) Modifying assembly conditions to minimize DNA damage leading to mutations, 3) Global error correction of the assembled product by biochemical methods directed at producing the consensus sequence of the assembled product, and 4) Efficient correction of errors by oligonucleotide-directed mutagenesis following cloning and sequencing of the assembled product.

To drive development of improved methodology we are undertaking some larger synthetic projects. We are assembling the complete genome of the mouse mitochondrion. We are also synthesizing a region from the *M. genitalium* genome that encodes genes essential for translation of mRNA to produce proteins.

* Presenting author

44

Development of a *Deinococcus radiodurans* Homologous Recombination System

Sanjay Vashee*, Ray-Yuan Chuang* (RChuang@venterinstitute.org), Christian Barnes, Hamilton O. Smith, and J. Craig Venter

J. Craig Venter Institute, Rockville, MD

A major goal of our Institute is to rationally design synthetic microorganisms that are capable of carrying out the required functions. One of the requirements for this effort entails the packaging of the designed pathways into a cohesive genome. Our approach to this problem is to develop an efficient in vitro homologous recombination system based upon *Deinococcus radiodurans* (Dr). This bacterium was selected because it has the remarkable ability to survive 15,000 Gy of ionizing radiation. In contrast, doses below 10 Gy are lethal to almost all other organisms. Although hundreds of double-strand breaks are created, Dr is able to accurately restore its genome without evidence of mutation within a few hours after exposure, suggesting that the bacterium has a very efficient repair mechanism. The major repair pathway is thought to be homologous recombination, mainly because Dr strains containing mutations in *recA*, the bacterial recombinase, are sensitive to ionizing radiation.

Since the mechanism of homologous recombination is not yet well understood in Dr, we have undertaken two general approaches to study this phenomenon. First, we are establishing an endogenous extract that contains homologous recombination activity. This extract can then be fractionated to isolate and purify all proteins that perform homologous recombination. We are also utilizing information from the sequenced genome. For example, homologues of *E. coli* homologous recombination proteins, such as recD and ruvA, are present in Dr. Thus, another approach is to assemble the homologous recombination activity by purifying and characterizing the analogous recombinant proteins. However, not all genes that play a major role in homologous recombination have been identified by annotation.

We have over-expressed and purified *D. radiodurans* homologues of repair proteins, including RecA, SSB, RecD, RuvB, and RuvC. We showed that the properties of DrSsb DNA binding and strand-exchange properties are very similar to that of *E. coli* SSB. In addition, using antibodies we have raised, we have determined that the amount of both DrSsb and RecA protein increases in the cell when exposed to a DNA damaging agent, whereas the level of DrRecD protein remains the same.

* Presenting author

45

Development of a Novel Recombinant Cyanobacterial System for Hydrogen Production from Water

Qing Xu, Shibu Yooseph, Hamilton O. Smith, and J. Craig Venter (jcventer@tcag.org)

J. Craig Venter Institute, Rockville, MD

Hydrogen is a clean alternative to gasoline and other fossil fuels as it generates only water as a by-product. The development of cost-effective and renewable approaches to produce hydrogen fuel will lead to a new and efficient energy system, which will address both the adverse environmental impacts of fossil fuels and the need for energy independence. Photobiological processes are attractive routes to renewable hydrogen production. With solar energy, photosynthetic microbes such as cyanobacteria can extract energy from water via oxygenic photosynthesis. The resulting energy can be coupled to a hydrogenase system that yields hydrogen. However, one major drawback of this process is that most hydrogen-evolving hydrogenases are irreversibly inhibited by oxygen, which is an inherent byproduct of oxygenic photosynthesis.

The overall goal of our project is to develop a novel, O_2-tolerant photobiological system in cyanobacteria that can produce hydrogen continuously using water as the substrate. We have undertaken two general approaches to achieve this goal. Our first approach is to transfer known O_2-tolerant NiFe-hydrogenase into cyanobacteria. It is reported that purple-sulfur photosynthetic bacterium *Thiocapsa roseopersicina* has an oxygen-tolerant evolving NiFe-hydrogenase, but this anoxygenic microbe can not split water. It is therefore logic to transfer the hydrogenase into cyanobacteria to construct a novel hybrid system that physically combines the most desirable properties of two bacteria. Thus far, we have cloned O_2-tolerant hydrogenase genes *hydS* and *hydL* from *T. roseopersicina* into a cyanobacterial expression vector to create plasmid pEX-Tr, which was then transformed into *Synechococcus sp.* PCC7942. Analysis of pEX-Tr *Synechococcus* transformants are currently in the process. Our second approach is to identify putative O_2-tolerant NiFe-hydrogenases from marine microbes and transfer them into cyanobacteria. Screening hydrogenase genes from the environment has been a useful approach to find novel O_2-tolerant hydrogenases. However, to date, only a few O_2-tolerant hydrogenases have been identified. To take advantage of the environmental genetic information generated by our ongoing Global Ocean Sampling Project, we currently use probabilistic modeling methods such as HMMs to search our sequence databases for putative NiFe-hydrogenases. We will transfer them into cyanobacteria for heterologous expression, and then screen for novel O_2-tolerant hydrogenases using a chemochromic screening method developed by scientists at NREL. Thus far, we have identified a putative NiFe-hydrogenase that shows strong homology to a known O_2-tolerant hydrogenase. Cloning the genes is in the process.

46

Biotechnology For the Production of Ethanol and Butanol from Cellulose

Prabha P. Iyer* (piyer@venterinstitute.org), Hamilton O. Smith, and J. Craig Venter

J. Craig Venter Institute, Rockville, MD

Cellulose is the largest component of all biomass and is the most abundant organic polymer in nature. It also has the potential to be fermented into a number of useful products including ethanol and butanol. This has important implications for the production of renewable, biomass-based fuels. However, the crystalline structure of cellulose makes it a challenging substrate for most bacteria to break down. Further, most of the bacteria that do degrade cellulose are not able to produce large amounts of ethanol or butanol. The goal of this project is to construct a microorganism which can break cellulose down to glucose monomers and then ferment the glucose to appreciable quantities of alcohols. This would eliminate the current requirement for large amounts of enzymes to depolymerize cellulose before it can be fermented to alcohols.

We are in the process of cloning and heterologously expressing the entire cellulosome gene cluster of *Clostridium cellulolyticum* in *Clostridium acetobutylicum*. *C. cellulolyticum* is able to degrade crystalline cellulose; however, it is not able to produce large amounts of ethanol. On the other hand, *C. acetobutylicum* is a very well characterized organism which has traditionally been used to produce solvents in acetone-butanol-ethanol fermentations. Production of a cellulosome in *C. acetobutylicum* would constitute a major advance towards consolidated bioprocessing of cellulose to ethanol.

Expression of the cellulosome in organisms capable of producing butyrate such as *Clostridium butyricum* and *Clostridium tyrobutyricum* is also being explored. This would facilitate the degradation of cellulose to butyrate which can then be converted into butanol via a second fermentation reaction.

* Presenting author

Communication

47

Communicating Genomics:GTL

Anne E. Adamson, Shirley H. Andrews, Jennifer L. Bownas, Denise K. Casey, Sherry A. Estes, Sheryl A. Martin, Marissa D. Mills, Kim Nylander, Judy M. Wyrick, Anita J. Alton, and Betty K. Mansfield* (mansfieldbk@ornl.gov)

Oak Ridge National Laboratory, Oak Ridge, TN

Project Web site: DOEGenomeToLife.org

To help build the critical multidisciplinary research community needed to advance systems microbiology research, the Genome Management Information System (GMIS) contributes to DOE Genomics:GTL program strategies. GMIS also communicates key scientific and technical concepts emanating from GTL and related programs to the scientific community and the public. We welcome ideas for extending and improving communications and program integration to represent GTL science more comprehensively to multidisciplinary audiences.

Accelerating GTL Science

For the past 16 years, we have focused on presenting Human Genome Project (HGP) information and on imparting knowledge to a wide variety of audiences. Our goal always has always been to help ensure that investigators could participate in and reap the scientific bounty of the genomic revolution, new generations of students could be trained, and the public could make informed decisions regarding complicated genetics issues. Since 2000, GMIS has built on this experience to communicate about the DOE Office of Science's Genomics:GTL program, sponsored jointly by the Office of Biological and Environmental Research and the Office of Advanced Scientific Computing Research.

GTL is a departure into a new territory of complexity and opportunity requiring contributions of interdisciplinary teams from the life, physical, and computing sciences and necessitating an unprecedented integrative communications approach. Because each discipline has its own perspective, effective communication is highly critical to the overall coordination and success of GTL. Part of the challenge is to help groups speak the same language, from team and research-community building and strategy development through program implementation and reporting of results to technical and lay audiences. Our mission is to inform and foster participation by the greater scientific community and administrators, educators, students, and the general public.

Specifically, our goals center on accelerating GTL science and subsequent applications. They include the following:

- Foster information sharing, strategy development, and communication among scientists and across disciplines to accomplish synergies, innovation, and increased integration of knowledge. A new research community centered around the advanced concepts in GTL will emerge from this effort.
- Help reduce duplication of work.

* Presenting author

- Increase public awareness about the importance of understanding microbial systems and their capabilities. This information is critical not only to DOE missions in energy and environment but to the international community as well.

For the past year we have been working with DOE staff and teams of scientists to develop the next program and facilities roadmap for GTL. This roadmap, a planning and program-management tool, will be reviewed by the National Academy of Sciences. Tasks have included helping to organize workshops, capture workshop output for the roadmap, and conduct the myriad activities associated with creating a technical document of the roadmap's size and importance.

In the last 12 months, we also overhauled the GTL website for the growing program. All GTL publications are on the public Web site, which includes an image gallery, research abstracts, and links to program funding announcements and individual researcher Web sites. Additional site enhancements will be implemented once the GTL roadmap is published.

In addition to the GTL Web site, we produce such related sites as Human Genome Project Information, Microbial Genome Program, Microbial Genomics Gateway, Gene Gateway, Chromosome Launchpad, and the CERN Library on Genetics. Collectively, HGMIS Web sites receive more than 15 million hits per month. Over a million text-file hits from more than 300,000 user sessions last about 13 minutes—well above the average time for Web visits. We are leveraging this Web activity to increase visibility for the GTL program.

For outreach and to increase program input and grantee base, we identify venues for special GTL symposia and presentations by program managers and grantees. We present the GTL program via our exhibit at meetings of such organizations as the American Association for the Advancement of Science, American Society for Microbiology, American Chemical Society, National Science Teachers Association, National Association for Biology Teachers, and Biotechnology Industry Organization. We mail some 1600 packages of educational material each month to requestors and furnish handouts in bulk to meeting organizers who are hosting genomics educational events. We continue to create and update handouts, including a primer that explores the impact of genomics on science and society and flyers on careers in genetics and on relevant issues of concern to minority communities. We supply educational materials in print and on the Web site about ethical, legal, and social issues (called ELSI) surrounding the increased availability of genetic information.

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

* Presenting author

Bioinformatics, Modeling, and Computation

48

SimPheny™: A Computational Infrastructure for Systems Biology

Christophe H. Schilling* (cschilling@genomatica.com), Sean Kane, Martin Roth, Jin Ruan, Kurt Stadsklev, Rajendra Thakar, Evelyn Travnik, Steve van Dien, and Sharon Wiback

Genomatica, Inc., San Diego, CA

The Genomics:GTL (GTL) program has clearly stated a number of overall goals that will only be achieved if we develop "a computational infrastructure for systems biology that enables the development of computational models for complex biological systems that can predict the behavior of these complex systems and their responses to the environment." At Genomatica we have developed the SimPheny™ (for Simulating Phenotypes) platform as the computational infrastructure to support a model-driven systems biology research paradigm. SimPheny enables the efficient development of genome-scale metabolic models of microbial organisms and their simulation using a constraint-based modeling approach.

With SBIR grants from the DOE Genomics:GTL program we are in the process of enhancing the system-level functionality within SimPheny based on a collection of well qualified input from our pilot users. We have successfully deployed SimPheny into a series of academic laboratories and addressed key issues related to deployment strategies, collaboration requirements, training/support services, and experimental data integration needs, as well as modeling and simulation requirements. The findings that have resulted from these efforts have led to the identification of a series of high priority system-level requirements that need to be incorporated into the SimPheny platform in order to facilitate the use of such technologies to establish a model-driven research paradigm that can be broadly utilized by both expert and non-expert users.

We are now focusing on the software research and development activities necessary to achieve these system requirements. They include a major rework of the SimPheny underlying architecture and the related addition of functionality to improve remote connectivity for collaboration. Furthermore we are developing functionality to enable model and simulation data to be exported and imported within SimPheny. We are also developing capabilities to enable model comparison and the publishing of model content.

An integrated Fermentation Module has recently been added to SimPheny. This module allows users to manage, analyze, visualize, and integrate fermentation data within the existing modeling infrastructure. The module has the capability to check for the consistency of fermentation data by conducting elemental material balances and to convert raw concentration measurements into flux data that can be used to constrain and further validate simulation studies within SimPheny.

We are also in the process of integrating flux measurements within the context of genome-scale metabolic models to further our ability to analyze, interpret, and predict the behavior of biological systems. Designs are currently under way to incorporate data from 13C-labeling experiments into SimPheny and interpret the resulting flux measurements in the context of the predictive genome-scale models. Such functionality in SimPheny will provide experimental verification for our constraint-based

* Presenting author

predictions of metabolic flux distributions as well as lead to a streamlined method for isotope label tracing experiment analysis that can be implemented by the non-expert user.

These enhancements and overall system improvements will create an extremely stable foundation that will enable SimPheny to be deployed to a broader range of users with more targeted functionality. These efforts will lead to the creation of a comprehensive package of software, model content, and training/support for the delivery of cellular modeling technologies to the metabolic research community.

49

Hybrid Bacterial Cell Models: Linking Genomics to Physiological Response

Jordan C. Atlas[1]* (jca33@cornell.edu), Mariajose Castellanos[1], Anjali Dhiman[1,2], Bruce Church[2], and Michael L. Shuler[1]

[1]Cornell University, Ithaca, NY and [2]Gene Network Sciences, Ithaca, NY

A major challenge in the biological sciences is to relate cell physiology to genomic structure. Models that explicitly link genomic and proteomic data to physiology are necessary to take full advantage of bioinformatics. We have developed a whole cell hybrid model that captures the dynamics of a single celled chemoheterotrophic prokaryote. By hybrid model we refer to inserting a genomically/molecularly detailed sub-model into a coarse-grained model which is embedded in a representation of the cell's environment. The initial step is to construct a coarse-grained model with lumped "pseudochemical species" (lumped components of similar chemical species). All subsystems of the coarse-grained model can be "de-lumped" into genomically complete, chemically distinct subsystems with corresponding genes and gene products. Using this coarse-grained host model structure we expect to quickly build a complete coarse-grained model of any given chemoheterotrophic bacteria using data from chemostat or other growth experiments. By combining molecularly detailed modules within the coarse grained host model, we capture not only the internal details of the dynamics of the molecular subsystem, but also can evaluate that mechanism within the context of a whole cell and its environment. The whole cell modeling approach presented here is being augmented by statistical mechanics methods for parameter estimation that allow us to rapidly develop parameter sets for new modules as they are added.

This framework has been applied to create Cornell's Minimal Cell Model (MCM). The MCM is a theoretical construction that attempts to develop our understanding of the relationship between cellular function and genetics using a "bottom up" approach; the necessary model functions are selected by rationally deciding what machinery a cell needs to live and reproduce. A "minimal cell" is a hypothetical free living cell possessing the functions required for sustained growth and reproduction in a maximally supportive culture environment. The MCM simulates the growth of a minimal cell. Ultimately, we aim to model the complete functionality of a minimal cell. The Shuler group has demonstrated the "modularity" of hybrid models by constructing a genomically and chemically detailed model of nucleotide metabolism within the MCM (PNAS v. 101(17), pp. 6681-6686). The current work focuses on incorporating amino acid supplementation into the coarse grained model. Another system of interest is *Shewanella oneidensis*, which has the potential to help remove metal pollutants from the environment. We believe that these techniques will ultimately allow us to build a model for *Shewanella* that creates a connection from the organism's genomics, to its molecular functions, to the whole cell, and to the environment.

* Presenting author

50

Identification of the Most Probable Biological Network Using Model Discrimination Analysis

Andrea L. Knorr and Ranjan Srivastava* (srivasta@engr.uconn.edu)

University of Connecticut, Storrs, CT

In seeking to understand the behavior of biological systems, whether at the molecular, cellular, or higher level, it is possible to develop multiple hypotheses of how the system of interest functions. These hypotheses may often be formulated into different network descriptions of the system. Using a Bayesian-based model discrimination technique, it is possible to determine which network, and as a consequence, which hypothesis is most probable. It is important to note that this method of network determination is not a data-mining approach, but rather is hypothesis-driven.

As an illustration of model discrimination, our work evaluating and identifying the most probable model of HIV-1 viral dynamics will be presented. Four different models of viral dynamics accounting for uninfected cells, infected cells, viral level, and/or the immune response were either taken from the literature or developed by our group. Parameters for the models were estimated from a cohort of 338 patients monitored for up to 2,484 days. Model discrimination was applied to determine which of the models, based on how they best captured overall viral dynamics, was most probable. The model determined as most likely was overwhelmingly favored relative to the remaining three models. It accounted for uninfected cells, infected cells, and cytotoxic T lymphocyte dynamics. Interestingly it was the only model that did not explicitly account for viral load, suggesting that none of the models to date have captured the appropriate network connectivity relating to viral load.

The technique of model discrimination is generic enough that it may easily be used to analyze biochemical kinetic pathways or identify the most likely genetic regulatory network in a given system of interest. To make such analysis readily available to a larger user base, we are in the process of developing a software package for carrying out model discrimination. Specifically, the package will allow the user to enter their models using SBML, as well as the appropriate data. The package will then determine the most probable model, presenting statistical analysis and comparative graphical output of actual and predicted network behavior.

* Presenting author

51

Rhodopseudomonas palustris Regulons Detected by a Cross-Species Analysis of the α-Proteobacteria

Sean Conlan[1,*] (sconlan@wadsworth.org), Charles E. Lawrence[1,2], and Lee Ann McCue[1]

[1]The Wadsworth Center, Albany, NY and [2]Brown University, Providence, RI

The objective of this study is to elucidate transcription regulatory mechanisms of the environmentally significant bacterium, *Rhodopseudomonas palustris*. This α-proteobacterial species carries out three of the chemical reactions that support life on this planet: the conversion of sunlight to chemical-potential energy, the absorption of carbon dioxide which it converts to cellular material, and the fixation of atmospheric nitrogen into ammonia. We predicted regulatory signals genome-wide by applying a Gibbs sampling algorithm[1,2,3] to orthologous intergenic regions; specifically, those upstream of 2,044 genes/operons from *R. palustris* and seven other α-proteobacterial species (*Bradyrhizobium japonicum*, *Brucella suis*, *Caulobacter crescentus*, *Rhodobacter sphaeroides*, *Rhodospirillum rubrum*, *Hyphomonas neptunium*, and *Novosphingobium aromaticivorans*). A Bayesian motif clustering algorithm[4] was then used to cluster the cross-species motifs to identify genes that are likely regulated by the same transcription factor (i.e., a regulon). Of the 101 putative regulons detected, several were of particular interest: an organic hydroperoxide resistance regulon, a flagellar regulon, a photosynthetic regulon, the LexA regulon, and four regulons involved in nitrogen metabolism (FixK$_2$, NnrR, NtrC, σ^{54}). In addition, a highly conserved repeat sequence was detected downstream of over 100 genes.

Cognate transcription factor identification: Organic hydroperoxide resistance

At the core of the transcription regulatory network is the interaction between a transcription factor and its cognate binding site. Currently, there is no reliable way to infer these *cis-trans* connections from sequence data alone. It has been estimated, however, that ~50% of transcription factors in *E. coli* are auto-regulatory. Given that, auto-regulatory site(s) for a transcription factor should cluster with sites for additional genes that are regulated. We investigated eight *R. palustris* clusters (mentioned above) with biochemical and genetic data in the literature and found that four of those clusters contain motifs upstream of the likely cognate transcription factor. In particular, a motif upstream of *rpa0828*, a MarR family transcription factor of unknown specificity, clustered with motifs upstream of two genes involved in resistance to organic hydroperoxides (*ohr*, *rpa4101*). RPA0828, and its orthologs, all contain a highly-conserved cysteine residue required for the activity of the organic hydroperoxide resistance regulator, OhrR, from *Xanthamonas campestris* (Fig. 1). Therefore, it seems likely that RPA0828 is a regulator of hydroperoxide resistance in *R. palustris*. An additional OhrR homolog, RPA4102, was found in the *rpa4101-4103* operon and may have a similar or redundant function.

Figure 1. Alignment of the oxidation sensitive regions of several OhrR orthologs.

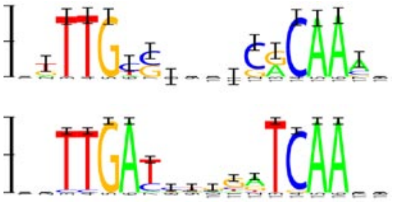
Organism	Protein	Alignment
X. campestris	OhrR	LDNQL**C**FALYS
R. palustris	RPA0828	LETQL**C**FALYS
R. palustris	RPA4102	LDRQV**C**FLLYA
B. japonicum	BLR0736	LDNQI**C**FAVYS
B. suis	BRA0886	LADML**C**FAVYS
H. neptunium	n.d.	LDHAL**C**FAIYS

* Presenting author

Discriminating between members of a transcription factor family: FixK$_2$ and NnrR

A difficulty with any clustering procedure is determining how many clusters are present in the data set, and many clustering approaches require this knowledge *a priori*. The Bayesian motif clustering algorithm (BMC), used in our work, determines the number of clusters based on sequence evidence and a tunable parameter (q), which influences whether a motif forms a cluster by itself or joins an existing cluster. This ability of the BMC algorithm to infer the number of clusters, while detecting subtle differences between motif types, is demonstrated by the FixK$_2$ and NnrR clusters. FixK$_2$ and NnrR regulate genes involved in respiration and nitric oxide metabolism, respectively, and both belong to the Fnr/Crp transcription factor family. Genes involved in these two pathways (respiration and nitric oxide metabolism) formed distinct clusters with the logos shown in Figure 2. Despite the high similarity between the binding consensus sequences, and 55% identity between the helix-turn-helix regions of FixK$_2$ and NnrR, BMC correctly separated the motifs of these regulons.

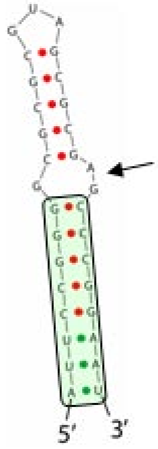
Figure 2. Sequence logos of the FixK$_2$ (top) and NnrR (bottom) clusters.



Detection of a novel inverted repeat

A benefit of the genome-wide approach used in this study is that conserved regulatory signals, regardless of their mechanism, can be detected. By not limiting the search to a particular set of genes (*e.g.*, a set of genes identified by a microarray experiment), or to a particular transcription factor, we were able to find a highly conserved inverted repeat downstream of over 100 genes. This repeat was found almost exclusively in intergenic regions at the 3' ends of genes. It was very rarely found between divergently transcribed genes or in coding regions. It is composed of a variable region, flanked by invariant inverted repeats. The variable region, which is 10-52 bp in length, is palindromic in 89% of the cases. Analysis of the repeats using *S*fold[5], demonstrated that many likely fold into a structure composed of an invariant helix, followed by a bulge and a variable-length hairpin (Fig. 3). While the repeat has features reminiscent of a mobile DNA element or transcriptional terminator, neither of these elements provide satisfactory explanations for the non-random distribution and perfectly conserved ends.

Figure 3. Putative structure of a 32 bp inverted repeat. The invariant helix is boxed and the bulge is indicated with an arrow.



References

1. Thompson, W, Rouchka, EC and Lawrence, CE. *NAR* **31**:566-84 (2003).
2. McCue, LA, Thompson, W, Carmack, CS, Ryan, JS, Liu, JS, Derbyshire, V and Lawrence, CE. *NAR* **29**: 774-782 (2001).
3. McCue, LA, Thompson, W, Carmack, CS and Lawrence, CE. *GenRes* **12**: 1523-32 (2002)
4. Qin, ZS, McCue, LA, Thompson, W, Mayerhofer, L, Lawrence, CE and Liu, JS. *NatBiotech* **21**: 435-39 (2003).
5. Ding, Y, Chan, CY and Lawrence, CE. *NAR* **32**: W135-41 (2004).

* Presenting author

52

Exploring Evolutionary Space

Timothy G. Lilburn[1]* (tlilburn@atcc.org), Yun Bai[2], Yuan Zhang[2], James R. Cole[2], and George M. Garrity[2]

[1]American Type Culture Collection, Manassas, VA and [2]Michigan State University, East Lansing, MI

The use of principal components analysis (PCA) to visualize the evolutionary relationships among thousands of sequences was developed by us as a tool for aiding in the definition of higher-level prokaryotic taxonomy. It not only helped define higher taxa by revealing naturally occurring clusters within the data, but also proved invaluable in highlighting errors in classification and annotation of sequences. Such PCA projections can be viewed as maps of evolutionary space for the sequences and, by extension, the organisms (and genomes) from which the sequences are obtained. Maps based on SSU rRNA sequences show large gaps between some phylogenetic groups. Presumably, this white space is due to the constraints on the evolution of these molecules that arise from their functional requirements. Sequences that might appear there either simply cannot occur in nature or belong to extinct species. Although PCA and other projection techniques can provide a reasonable approximation of the topology hidden within a dataset, some distortion is inevitable and can be attributed to methodological biases and biases that may exist within the data. Previously, we had demonstrated that we could improve the accuracy of projections for a test case having a known topology and coordinate system by using a set of uniformly distributed external benchmarks. However, neither the true topology nor the coordinate system of the prokaryotic evolutionary space has been defined. Therefore, to understand the distortion, we would need to first define the limits of this space. In this study, we examine the use of a limited number (n=179) of internal reference points (benchmarks) on the transformation of the evolutionary distance data into the new coordinate system defined by PCA. We look at ways of making our maps of evolutionary space more accurate and explore why the white space exists. Methods explored include the generation of synthetic polychimeras, *in silico* random mutation, and complementation of a set of 179 proposed benchmark sequences. Our results are presented as a set of PCA plots that are evaluated in terms of their resolution and their concordance with the current taxonomy and with each other.

53

PhyloScan: a New Tool for Identifying Statistically Significant Transcription Factor Binding Sites by Combining Cross-Species Evidence

Lee A. Newberg[1,2]*, C. Steven Carmack[1], Lee Ann McCue[1] (mccue@wadsworth.org), and Charles E. Lawrence[3]

[1]Wadsworth Center, Albany, NY; [2]Rensselaer Polytechnic Institute, Troy, NY; and [3]Brown University, Providence, RI

If there are known transcription factor binding sites (TFBSs) for a particular transcription factor (TF), then it is possible to construct a motif model or position weight matrix with which to scan a

* Presenting author

sequence database for additional sites, thereby predicting a regulon. However, scanning a genome for additional TFBSs typically results in finding few statistically significant sites. Specifically, the statistical significance of a sequence match (p-value) to a motif can be assessed by comparison with the probability of observing a match with a score as good or better in a randomly generated search space of identical size and nucleotide composition -- the smaller the p-value the greater the evidence that the match is not due to chance alone. Staden [1] presented an efficient method that exactly calculates this probability, and Neuwald *et al*. [2] described an implementation of this method. In practice, when scanning a genome or the promoter regions of a genome, it is frequently difficult to identify (below a chosen level of statistical significance) even the known TFBSs that were used in the construction of the motif, to say nothing of additional novel sites for that TF. Essentially, given the statistical nature of this approach, only a relatively small number of TFBSs will be identified that could possibly be considered significant (low sensitivity, high specificity).

With the goal of increasing the statistical power of scanning a genome sequence database with a regulatory motif, we have developed a scanning algorithm, PhyloScan, that combines evidence from matching sites found in orthologous data from several related species. Specifically, we have extended Staden's method [1] to allow scanning of orthologous sequence data that is either multiply-aligned, unaligned or a combination thereof (aligned and unaligned). PhyloScan statistically accounts for the phylogenetic dependence of the species contributing aligned data and returns a p-value for the sequence match; importantly, the statistical significance is calculated directly, without employing training sets.

To evaluate this method we chose the *Escherichia coli* Crp and PurR motifs and gathered genome sequence data for several gamma-proteobacteria. Among the species chosen for this study (*E. coli*, *Salmonella enterica* Typhi, *Yersinia pestis*, *Haemophilus influenzae*, *Vibrio cholerae*, *Shewanella oneidensis*, and *Pseudomonas aeruginosa*), only *E. coli* and *S. typhi* exhibit extensive homology in the promoter regions [3]. Thus we aligned orthologous intergenic regions for these two species, and combine statistical evidence from scanning the aligned *E. coli* and *S. typhi* data with statistical evidence from scanning unaligned orthologous intergenic regions from the remaining five more distantly related species. This method enhances the identification of TFBSs in *E. coli* by several-fold over scanning the set of *E. coli* intergenic regions alone.

1. Staden, R. (1989) Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci* **5:**89-96.
2. Neuwald, A. F., Liu, J. S. and Lawrence, C. E. (1995) Gibbs motif sampling: Detection of bacterial outer membrane protein repeats. *Protein Sci* **4:**1618-32.
3. McCue, L. A., Thompson, W., Carmack, C. S. and Lawrence, C. E. (2002) Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res* **12:**1523-32.

54

Predicting Protein Interactions via Docking Mesh Evaluator

Roummel F. Marcia[1], Susan D. Lindsey[2], Erick A. Butzlaff[1], and Julie C. Mitchell[1]* (mitchell@math.wisc.edu)

[1]University of Wisconsin, Madison, WI and [2]University of California, San Diego, CA

Introduction

The Docking Mesh Evaluator (DoME) is a software package for protein docking and energy evaluation. It uses fast energy evaluation methods and optimization algorithms to predict the docked configurations of proteins with DNA, ligands, and other macromolecules. A fully parallelized package, DoME can run on supercomputers, clusters, and linked independent workstations.

Description

Previously, DoME's energy model was based on solvent effects defined implicitly using adaptive mesh solutions to the Poisson-Boltzmann equation and van der Waals energy terms. While this earlier version achieved moderate success [1], more accurate results were obtained by incorporating hydrogen bond and solvation energy terms. The hydrogen bond term uses the spatial relationship between a potential acceptor atom and a donor-hydrogen pair both to determine the existence of a hydrogen bond and to compute the bond's potential energy. The solvation term uses a modification to the method of Zhang et al. [2] to estimate the effective atomic contact energy of a complex in water. In addition, the energy function employs weighting and switching parameters to measure the individual contribution of each energy term to the overall interaction.

Global optimization methods were developed to determine the lowest function value of this energy model. In particular, the General Convex Quadratic Approximation [3] constructs a sequence of convex underestimators to a collection of local minima to predict possible areas of low energies. Yukawa potentials are used as analytic solutions to the linearized Poisson-Boltzmann equations so that gradient-based local optimization can be performed. Initial scanning of the energy landscape for favorable configurations as initial seeds for underestimation improves algorithm performance. This coupled use of scanning and optimizing is more effective in determining points of low energy values than scanning or optimizing alone [4].

We present results from the standard benchmarking test set of Chen et al. [5] for testing protein-protein docking algorithms. We consider both bound and unbound crystalline structures for determining proper docked configurations. We also highlight which energy terms are significant in each test case. (Of the 59 the benchmarking test set contains, 22 are enzyme-inhibitor complexes, 19 are antibody-antigen complexes, 11 are various complexes, and 7 are difficult test cases whose solutions have considerable structural changes.)

References

1. R.F. Marcia, J.C. Mitchell, and J.B. Rosen, Iterative convex quadratic approximation for global optimization in protein docking, *Comput. Optim. Appl.*, Accepted for publication.
2. C. Zhang, G. Vasmatzis, J.L. Cornette, C. DeLisi, Determination of atomic desolvation energies from the structures of crystallized proteins, *J. Mol. Biol*, 1997 Apr. 4; 267(3); 707-26.

* Presenting author

3. J.B. Rosen and R.F. Marcia, Convex quadratic approximation, *Comput. Optim. Appl.*, 28, pp.173-184, 2004.
4. J.C. Mitchell, J.B. Rosen, A.T. Phillips, and L.F. Ten Eyck, Coupled optimization in protein docking, in *Proceedings of the Third Annual International Conference on Computational Molecular Biology*, ACM Press, 1999, pp. 280-284.
5. R. Chen, J. Mintseris, J. Janin, and Z. Weng, A protein-protein docking benchmark, *Prot. Struct. Fun. Gen.*, 52, pp. 88-91, 2003.

55

UC Merced Center for Computational Biology

Michael Colvin[1]* (mcolvin@ucmerced.edu), Arnold Kim[1], and Felice Lightstone[2]

[1]University of California, Merced, CA and [2]Lawrence Livermore National Laboratory, Livermore, CA

We are establishing a Center for Computational Biology (CCB) at the newest campus of the University of California. The CCB will sponsor multidisciplinary scientific projects in which biological understanding is guided by computational modeling. The center will also facilitate the development and dissemination of undergraduate and graduate course materials based on the latest research in computational biology. The Center is starting a number of activities that aim to recast biology as an information science:

1. Host multidisciplinary research projects in computational and mathematical biology that will provide a rich environment for graduate and undergraduate research.
2. Develop new mathematical and computational methods that are widely applicable to predictive modeling in the life sciences.
3. Develop and disseminate computational biology course materials that translate new research results into educational resources.
4. Extend the successes in achieving these objectives to other universities and "university-feeder" institutions such as community colleges.

This project is a multi-institutional collaboration including the new University of California campus at Merced, Rice University, Rensselaer Polytechnic Institute, and Lawrence Livermore National Laboratory, as well as individual collaborators at other sites. The CCB will foster a number of research projects that emphasize the role of predictive simulations in guiding biological understanding by funding graduate students and post-doctoral fellows with backgrounds in the mathematical and computational sciences to work on collaborative biology projects. Additionally, the center will work to translate this research into educational materials. UC Merced, as the first U.S. research university of the 21st century, offers many advantages for this new center, including an ideal venue for developing and implementing new courses and degree programs, a highly multidisciplinary faculty and organizational structure, and a strong commitment to educational outreach to diverse and underrepresented groups. New computational biology courses will be used and assessed in the new UC Merced Biological Sciences major that will be accepting freshman and junior transfers in Fall 2005. Graduate courses will be implemented in the Quantitative Systems Biology Graduate Group that began accepting graduate students in Fall 2004. All course materials will be released under an open

* Presenting author

public license using the Connexions courseware system developed at Rice University. We anticipate that this new biology curriculum will be effective in attracting students to biology who have an interest and aptitude in mathematics and computational sciences, as well as broaden the horizons of students expecting a traditional biology program. The electronic, modular course materials produced by the center will facilitate linkages to feeder schools at the state university, community college, and high school levels.

The long-term impact of the CCB will be to help train a new generation of biologists who bridge the gap between the computational and life sciences and to implement a new biology curriculum that can both influence and be adopted by other universities. Such scientists will be critical to the success of new approaches to biology, exemplified by the DOE Genomics:GTL program in which comprehensive datasets will be assembled with the goal of enabling predictive modeling of the behavior of microbes and microbial communities, as well as the biological components of life, such as multiprotein machines.

56

Biomic Approach to Predictive Cell Modeling

P. J. Ortoleva* (ortoleva@indiana.edu), L. Ensman, J. Fan, K. Hubbard, A. Sayyed-Ahmad, F. Stanley, K. Tuncay, and K. Varala

Indiana University, Bloomington, IN

Replication, transcription, translation and metabolism, as well as physiological dynamics, are all strongly coupled and must be accounted for if predictive cell modeling is to be attained. Key barriers to doing so are that many processes are not yet understood in detail, and that the phenomenological parameters in a kinetic cell model are not yet calibrated. In this presentation we show how multiplex data (notably cDNA microarray time series) and incomplete cell models can be integrated via information theory to overcome these difficulties.

A kinetic cell model is expressed as $\frac{\partial \Psi}{\partial t} = F\left[\Psi, \Lambda, \Phi\right]$ (1)

where Ψ is a set of descriptive variables (RNA populations, metabolite, concentrations, etc.), Λ is a set of model parameters and Φ is a set of cell descriptive variables for which we do not have governing equations (e.g. concentrations of species for which reactions have not yet been delineated). The problem is that (1) we cannot run such an incomplete model as the timecourse $\Phi(t)$ is required; and (2) therefore we cannot calibrate the model parameters Λ.

To overcome these difficulties, in our procedure we use information theory to construct the probability ρ that is a function of Λ and a functional of the timecourse $\Phi(t)$. We use available data (notably cDNA microarray, proteomics, NMR) in steadv-state or time series to construct ρ. With this we seek the most probable values of $\Lambda, \Phi(t)$ by solving $\partial\rho/\partial\Lambda = 0, \delta\rho/\delta\Phi(t) = 0$, the latter being a functional differential equation for the timecourse $\Phi(t)$. Regularization is also used to insure that there are no unphysically short timescale effects in $\Phi(t)$ created by the sparseness of, or noise in, the data.

The above methodology has been implemented for the case of transcription/translation modeling integrated with cDNA microarray time series analysis. The input of our software is microarray time series and a putative transcription control network (the latter being incomplete and possibly error-

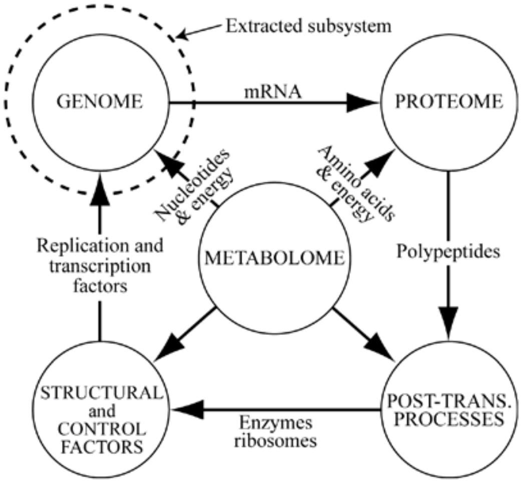
* Presenting author

prone). Output is calibrated transcription factor (TF) binding constants and transcription rate coefficients for all genes. Degradation rate constants for each type of RNA are also provided as are the timecourse of TF thermodynamic activities.

Our methodology has been tested on *E. coli* where we have identified some likely mistakes in existing *E. coli* networks and have delineated the TF timecourse that coordinate the change in transcription patterns accompanying a change in carbon source in the surroundings. The method is found to be very robust to omnipresent noise in the microarray data and to allow one to discover errors in the proposed regulatory network or to discover new TF/gene interactions. Possible interactions identified using sequence analysis are screened and their up/down regulatory function is identified.

The implications of our approach are far-reaching. The approach is applicable to large systems (thousands of genes, hundreds of transcription factors). Its multiplex and automated character will greatly accelerate the delineation of the gene regulatory network. The many rate and thermodynamic constants calibrated will make cell biomic modeling feasible. As suggested in Fig. 1, our approach allows for the piecewise development and calibration of a cell biomic model.

Figure 1. An extracted subsystem (here the genome) can be run and calibrated using probability functional information theory.



The hierarchical nature of the organization of intracellular structure and their multiple dimensional character are key to cell function. A cell must be understood in terms of its specialized zones wherein reaction and transport occur and molecules are exchanged among these zones. In summary, a cell is a very complex molecular processor that involves dynamic on fibrils (1D), membranes (2D) and with bulk medium (3D). The multiple dimensional character of intracellular dynamics is accounted for in CellX which accounts for reaction-transport dynamics along membranes embedded in bulk media, and the exchange among these zones via boundary conditions. Recent experimental studies suggest that MinC, MinD, and MinE proteins play a key role in the location of the Z-ring. The absence of Min dynamics results in location of Z-rings near the poles and imprecise cell division. MinC is an inhibitor to the formation of the Z-ring. MinC and MinD oscillations are observed to be in phase whereas MinE oscillation is coupled to MinC and MinE dynamics. Results for the autonomous localization of the division plane and the segregation of two daughter chromosomes will be presented as example applications of CellX.

57

The BioWarehouse System for Integration of Bioinformatics Databases

Tom Lee, Valerie Wagner, Yannick Pouliot, and Peter D. Karp* (pkarp@ai.sri.com)

SRI International, Menlo Park, CA

BioWarehouse [1] is an open-source toolkit for constructing bioinformatics database (DB) warehouses. It allows different users to integrate collections of DBs relevant to the problem at hand. BioWarehouse can integrate multiple public bioinformatics DBs into a common relational DB management system, facilitating a variety of DB integration tasks including comparative analysis and data mining. All data are loaded into a common schema to permit querying within a unified representation.

BioWarehouse currently supports the integration of Swiss-Prot, TrEMBL, ENZYME, KEGG, BioCyc, NCBI Taxonomy, CMR, and the microbial subset of Genbank. Loader tools implemented in the C and Java languages parse and load the preceding DBs into Oracle or MySQL instances of BioWarehouse.

The presentation will provide an overview of BioWarehouse goals, architecture, and implementation. The BioWarehouse schema supports the following bioinformatics datatypes: chemical compounds, biochemical reactions, metabolic pathways, proteins, genes, nucleic acid sequences, features on protein and nucleic-acid sequences, organism taxonomies, and controlled vocabularies.

BioWarehouse is in use by several bioinformatics projects. An SRI project is developing algorithms for predicting which genes within a sequenced genome code for missing enzymes within metabolic pathways predicted for that genome [2]. BioWarehouse fills several roles within that project: it is used to construct a complete and nonredundant dataset of sequenced enzymes by combining protein sequences from the UniProt and PIR DBs, and by removing from the resulting dataset those sequences that share a specified level of sequence similarity. Our current research involves extending the pathway hole filling algorithm with information from genome-context methods such as phylogenetic signatures, which are obtained from BioWarehouse thanks to the large all-against-all BLAST results stored within CMR. Another SRI project is comparing the data content of the EcoCyc and KEGG DBs using BioWarehouse to access the KEGG data in a computable form.

BioWarehouse is supported by the Department of Energy and by DARPA through the DARPA BioSPICE program for biological simulation.

1. BioWarehouse Home Page http://bioinformatics.ai.sri.com/biowarehouse/
2. Green, M.L. and Karp, P.D., "A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases," *BMC Bioinformatics* 5(1):76 2004 http://www.biomedcentral.com/1471-2105/5/76.

* Presenting author

58

Building Large Biological Dynamic Models of *Shewanella oneidensis* from Incomplete Data

Ravishankar R. Vallabhajosyula[1]*(rrao@kgi.edu), Sri Paladugu[1], Klaus Maier[2], and Herbert M. Sauro[1]

[1]Keck Graduate Institute, Claremont, CA and [2]University of Stuttgart, Stuttgart, Germany

The objective of this study is to investigate certain issues related to building large dynamic models of reaction networks. In particular, emphasis is placed on the performance of computational methods, both in terms of accuracy and computational time. In addition, we are also interested in methods for approximating reaction rate laws when kinetic information is not readily available.

To investigate these questions we have constructed a hybrid model that is a first attempt at a model of *Shewanella oneidensis MR-1* that includes Glycolysis and TCA cycle. Understanding how energy pathways function in *S. oneidensis* is very important to modelling the interaction of this organism with its environment.

The SBML files for Glycolysis and TCA Cycle for *Shewanella* are available from KEGG [1]. These were used to assist in the construction of a test model. However, there is a lack of kinetic information for the dynamics of the underlying metabolic reactions in the KEGG database. To overcome this problem, estimates were made by comparing data from similar pathways in *Escherichia Coli*. While we were able to adapt a kinetic model of glycolysis from work published by [2], the data for the TCA cycle was still lacking. A hybrid model using Linlog kinetics was constructed for the unknown reaction kinetics in the TCA Cycle.

Approximating Rate Laws

Linlog kinetics [3] were recently developed as an attractive alternative to the commonly used Michaelis-Menten like kinetics. The underlying theoretical framework for this approach is described using an example of a branched pathway in [4]. Linlog kinetics provides a very good approximation to Michaelis-Menten kinetics and requires fewer parameters. The equations for Linear, Power-Law and Linlog approximations to Michaelis-Menten kinetics, carried out about an operating effector concentration S_0 are shown in Fig. 1.

Fig. 1 Equations for Michaelis-Menten and other Approximation Methods

$$\text{Michaelis-Menten} \quad V = \frac{V_{max}S}{K_m+S}$$

$$\text{Linear Approximation} \quad V = mS + c, \qquad \text{where } m = \frac{V_{max}K_m}{(K_m+S_0)^2} \text{ and } c = \frac{V_{max}S_0}{(K_m+S_0)^2}$$

$$\text{Power-Law Approximation} \quad V = \frac{V_{max}S_0}{K_m+S_0}\left(\frac{S}{S_0}\right)^{\beta}, \qquad \text{where } \beta = \frac{K_m}{K_m+S_0}$$

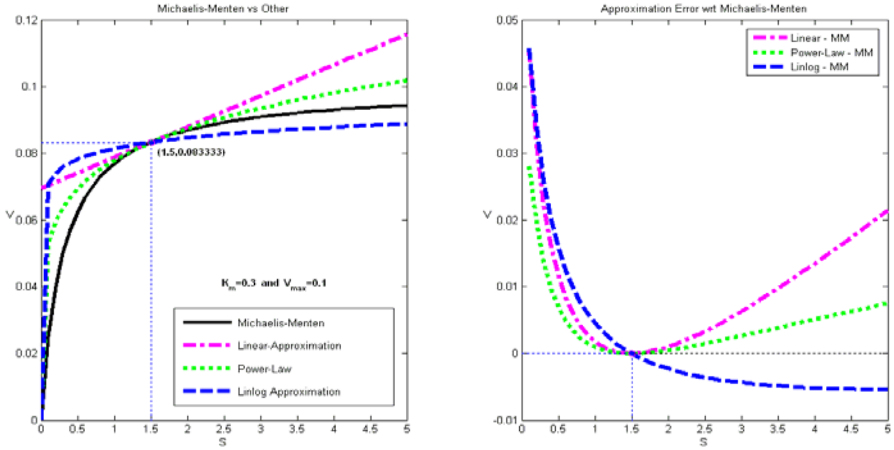
$$\text{Linlog Approximation} \quad V = \frac{V_{max}S_0}{K_m+S_0}.\left[1 + \epsilon\, ln\left(\frac{S}{S_0}\right)\right], \qquad \text{where } \epsilon = \frac{K_m}{K_m+S_0}$$

The equation describing the Linlog kinetics is a combination of a linear term with a logarithmic component involving the effector concentrations, scaled appropriately with the respective elasticities.

* Presenting author

A comparison of Linear, Power-law and Linlog approximations with Michaelis-Menten kinetics shows that Linlog has the least error over a wide range of effector concentrations. This comparison is shown in Fig. 2.

Fig. 2 Linear, Power-Law and Linlog Approximations vs. Michaelis-Menten kinetics



Performance Issues

The *Shewanella oneidensis* model was simulated using various software tools to investigate their performance given a complex network. These include publicly available simulators SCAMP and Jarnac, simulators built with languages such as FORTRAN, Java, C and C#, as well as simulators derived from commercially available software packages which included Mathematica and Matlab.

This study will provide useful insights into building a more powerful simulator that can handle especially complex networks necessary for carrying out large scale simulations. In this regard, it is essential to understand how existing simulators are structured internally. Since each of these simulators has a different approach to generating the solutions, one may be better than others at simulating a given network. For example, SCAMP and Jarnac generate optimized internal byte-code. Simulators for C and FORTRAN are based on compiled code. The C# and Java simulators are based on byte-code interpretation. The Java simulator code was generated in two ways, 1) The model equations were interpreted at run-time, and 2) A Java class was generated to function as a solver. Matlab performance was tested by generating a standard Matlab ODE function from the SBML of the *Shewanella* model. Mathematica code was similarly generated. All code other than Jarnac and SCAMP was generated using SBW SBML translator modules.

The results from the simulation of the combined Glycolysis and TCA cycle pathway in *Shewanella oneidensis* will be presented in the poster at the next GTL conference in 2005.

References

1. KEGG: Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg
2. C. Chassagnole, N. Noisommit-Rizzi, J.W. Schmid, K. Mauch, M. Reuss. "Dynamic Modeling of the central carbon metabolism of Escherichia Coli". *Biotechnol. Bioeng.*, **79**(1), pp.53-73, (2002)
3. L. Wu, W. Wang, W.A. van Winden, W.M. van Gulik and J.J. Heijnen. "A new framework for the estimation of control parameters in metabolic pathways using lin-log kinetics", *Eur. J. Biochem.*, **271**, pp.3348-3359, (2004).

* Presenting author

4. D. Visser, J.J. Heijnen. "Dynamic Simulation and metabolic re-design of a branched pathway using Linlog kinetics", *Metabolic Engineering*. **5**, pp.164-176, (2003).

59

A Bayesian Method for Identifying Missing Enzymes in Predicted Metabolic Pathway Databases

Michelle L. Green* (green@ai.sri.com) and Peter D. Karp

SRI International, Menlo Park, CA

The PathoLogic program constructs Pathway/Genome databases by using a genome's annotation to predict the set of metabolic pathways present in an organism. PathoLogic determines the set of reactions composing those pathways from the enzymes annotated in the organism's genome. Many enzymes in a genome may be missed during the initial annotation effort or may be assigned a non-specific function (e.g., "thiolase family protein"). These missing or incomplete annotations can result in *pathway holes*. Pathway holes occur when a genome appears to lack the enzymes needed to catalyze reactions in a pathway. If a protein has not been assigned a specific function during the annotation process, any reaction catalyzed by that protein will appear as a missing enzyme or pathway hole in a Pathway/Genome database.

We have developed a method [1] that efficiently combines homology and pathway-based evidence using Bayesian methods to identify candidates for filling pathway holes in Pathway/Genome databases. Our program, which is now part of the Pathway Tools software, identifies potential candidate sequences for pathway holes, and combines data from multiple, heterogeneous sources to assess the likelihood that a candidate has the required function. By considering not only evidence from homology searches, but also genomic and functional context (e.g., are there functionally-related genes nearby in the genome?), our algorithm emulates the manual sequence annotation process to determine the posterior belief that a candidate has the required function. The method can be applied across an entire metabolic pathway network and is generally applicable to any pathway database. We achieved 71% precision at a probability threshold of 0.9 during cross-validation using known reactions in computationally-predicted pathway/genome databases.

After applying our method to 255 pathway holes in 99 pathways from the CauloCyc database, the predictions from this program completed fourteen additional pathways. The program made putative assignments to 53 pathway holes, including annotation of 2 sequences of previously unknown function. The newly completed pathways include "fatty acid oxidation pathway", "oxidative branch of the pentose phosphate pathway", "peptidoglycan biosynthesis", "pyridine nucleotide biosynthesis", "pantothenate and coenzyme A biosynthesis", "de novo biosynthesis of pyrimidine ribonucleotides", "de novo biosynthesis of purine nucleotides II", "histidine biosynthesis I", "tyrosine biosynthesis I", "phenylalanine biosynthesis II", and "alanine biosynthesis I".

1. Green, M.L. and Karp, P.D., "A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases," *BMC Bioinformatics*, 5:76 2004. http://www.biomedcentral.com/1471-2105/5/76.

* Presenting author

60

Does EcoCyc or KEGG Provide a Preferable Gold Standard for Training and Evaluation of Genome-Context Methods?

Peter D. Karp* (pkarp@ai.sri.com) and Michelle L. Green

SRI International, Menlo Park, CA

Motivations: Genome-context methods such as phylogenetic profiles infer functional associations between genes and constitute a new approach to prediction of gene function. Most past developers of genome-context methods have trained and validated their algorithms against metabolic pathway DBs: primarily KEGG [1] and EcoCyc [2]. This work addresses the question of which of these two DBs is the optimal resource for training and evaluation of genome-context methods. Our hypothesis is that EcoCyc is the preferable resource to use because: (a) KEGG pathway maps contain many false-positive functional associations because KEGG maps tend to be much larger than EcoCyc pathways, and therefore contain genes from many different biological pathways, (b) KEGG pathways are less accurate because they are computationally predicted whereas EcoCyc pathways are curated from the biomedical literature, and (c) EcoCyc contains other types of functional relationships besides the metabolic and two-component signal transduction pathways that KEGG contains, such as descriptions of regulatory relationships between transcription factors and other genes.

Method: We evaluated this hypothesis by randomly choosing pairs of genes from the same KEGG and EcoCyc pathways, and counting the frequency with which those gene pairs show chromosomal adjacency, or similar phylogenetic profiles, since these basic methods for predicting functional associations are shown to have utility by both EcoCyc and KEGG.

Results: The hypothesis is validated by the following results. Two genes chosen at random from the same EcoCyc pathway were 4.7 times more likely to be adjacent on the chromosome than two genes chosen at random from the same KEGG map. Furthermore, gene pairs chosen from the same KEGG map that are not in the same EcoCyc pathway are even less likely to be chromosomally adjacent or to exhibit similar phylogenetic profiles. Similar results were obtained for the BioCyc and KEGG datasets for seven other organisms. In addition, two genes chosen at random from the same EcoCyc pathway were 3.0 times more likely to have similar phylogenetic profiles than two genes chosen at random from the same *E. coli* KEGG map. In addition, we find that transcription factors and the genes that they regulate show significant chromosomal adjacency and similar phylogenetic profiles.

Summary: EcoCyc and the BioCyc DBs for other organisms are preferable resources for training and evaluation of genome-context methods because their pathways are more biologically meaningful, and because they contain a wider range of biological functional associations, such as those between transcription factors and the genes they regulate, thus allowing genome-context methods to recognize a wider set of biological relationships.

1. M. Kanehisa, S. Goto, S. Kawashima and A. Nakaya, "The KEGG databases at GenomeNet," Nucleic Acids Research 40:42-6 2002.
2. I.M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I.T. Paulsen, M. Peralta-Gil, and P.D. Karp, EcoCyc: a comprehensive database resource for *E. coli, Nucleic Acids Research*, Database Issue, 2005 (in press).

* Presenting author

61

Towards a Physics and Systems Understanding of Ion Transport in Prokaryotes

Shreedhar Natarajan[1], Asba Tasneem*[1], Sameer Varma[1], Lakshminarayan Iyer[2], L. Aravind[2], and Eric Jakobsson*[2] (jake@ncsa.uiuc.edu)

[1]University of Illinois, Urbana, IL and [2]National Institutes of Health, Bethesda, MD

Ion transport mechanisms play three fundamental roles in biological systems: 1) generation and sensing of electrochemical signals, 2) generation of osmotic force for regulating water flow, and 3) energy transduction. It is useful to study these functions in an integrated manner because: 1) Ion transporters as they perform all three functions pose similar issues in understanding the physical bases of those functions, and 2) it is common for the same transporter to be critical for more than one function. Indeed, it is universally true that transporters which are critical for one function will affect the functioning of networks of transporters critical to other functions, because every transporter in a membrane affects osmotic, chemical, and electrical driving forces for every other transporter in the same membrane.

In this paper we report on our efforts along three lines: 1) By bioinformatics means we have discovered homologues of chemo-sensitive postsynaptic ion channels in prokaryotes, including in *Synechococcus,* which is the organism whose GTL project we are associated with. This discovery may point to a previously unknown molecular mechanism for electrochemical signaling between prokaryotes, and may also facilitate determination of structures for the ligand-gated channel family of gene proteins. 2) We have developed improved methods for assigning protonation states of electrically interacting titratable residues in the lumen of bacterial porins. This is critical in order to make realistic models for ion permeation through these channels. 3) We are adapting phylogenetic profiling methods to infer transport and regulatory networks that govern ion and water homeostasis in prokaryotes.

Supported by grant # 0235792 from the National Science Foundation, the U.S. Department of Energy's Genomics: GTL program (www.doegenomestolife.org) under project, "Carbon Sequestration in *Synechococcus Sp.*: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org), and the intramural program of the National Center for Biotechnology Information.

* Presenting author

62

OptStrain: A Computational Framework for Redesign Microbial Production Systems

Priti Pharkya and Costas D. Maranas* (costas@psu.edu)

Pennsylvania State University, University Park, PA

In this talk, we will discuss the hierarchical computational framework ***OptStrain*** aimed at guiding pathways modifications, through reaction additions and deletions, of microbial networks for the overproduction of targeted compounds. A comprehensive database of biotransformations, referred to as the Universal database (with over 5,700 reactions), is compiled and regularly updated by downloading and curating reactions from multiple biopathway database sources. Combinatorial optimization is then employed to elucidate the set(s) of non-native functionalities, extracted from this Universal database, to add to the examined production host for enabling the desired product formation. Subsequently, competing functionalities are identified and removed to ensure higher product yields coupled with growth. This work establishes an integrated computational framework capable of constructing stoichiometrically balanced pathways, imposing maximum product yield requirements, pinpointing the optimal substrate(s), and evaluating different microbial hosts. The range and utility of ***OptStrain*** is demonstrated by addressing a variety of product molecules and experimental verifications.

63

DEMSIM: A Discrete Event Based Mechanistic Simulation Platform for Gene Expression and Regulation Dynamics

Madhukar Dasika and Costas D. Maranas* (costas@psu.edu)

Pennsylvania State University, University Park, PA

The advent of high-throughput technologies has provided a major impetus for developing sophisticated computational frameworks to unravel the underlying regulatory circuitry that governs the response of biological systems to environmental and genetic perturbations. A systems engineering view reveals that gene expression dynamics are governed by processes that are essentially event driven, i.e., many events have to take place in a predetermined order with uncertain start and execution times to accomplish a certain task. There are many parallels between gene expression and manufacturing systems. In analogy to a manufacturing facility which produces a certain amount of finished product at a particular time with a certain probability, the transcription process produces mRNA transcripts with probability determined by the cellular environment and availability of required components. Similarly, accumulating mRNA and protein levels in the cell are akin to product inventory held in warehouses in a manufacturing system. Motivated by the numerous parallels between these two seemingly different settings, we have used discrete event simulation, which is a powerful tool employed to model and simulate supply chains and manufacturing systems, to model and simulate gene expression systems. In this talk, we will describe the DEMSIM tool that we have developed to test and hypothesize putative regulatory interactions.

* Presenting author

The key feature of the DEMSIM platform is the abstracting of underlying transcription, translation and decay processes as stand-alone modules. A module is further characterized by a sequence of discrete events. For example, the transcription module is composed of the following sequence of discrete events: (i) binding of RNA polymerase to promoter sequence (ii) transcription elongation and (iii) transcription termination. Each module is described by a set of physical and model parameters. Physical parameters correspond to parameters which are known *a priori* from literature sources and are fixed within the simulation framework (e.g. length of gene, transcription rate, etc.). In contrast, model parameters are regression parameters that are fitted using the available experimental data. Subsequently, the simulation is driven by the communication between these modules in accordance with the specifics of the regulatory circuitry of the biological system being investigated. The stochasticity inherent to all events is captured using Monte Carlo based sampling.

The DEMSIM software implementation consists of the following three key components: (i) an event list that contains all the events that need to be executed along with their respective execution times, (ii*)* a global simulation clock that records the progress of simulation time as events are sequentially executed, and (iii) a set of state variables that characterize the system and which are updated every time an event is executed. At every time step, events corresponding to all active (non–terminated) modules in the system are included in the event list. Subsequently, the event list is sorted and the event having the smallest execution time is executed. The simulation clock is advanced and the execution time of all other events is updated. Such a sequential procedure prevents the occurrence of causality errors by ensuring that an event with a later time stamps is not executed before an event with an earlier time stamp. Furthermore, since the execution of certain events leads to the creation of new modules and the termination of existing ones, the number of active modules in the system is updated and new events are included in the event list. This procedure is then repeated for the duration of the simulation horizon and state variables such as number of mRNA and protein molecules are recorded.

In this talk we will present the results for three biological systems of different levels of complexity that we have used to benchmark the DEMSIM platform. Simulation results for the relatively simple *lac* operon system of *E. coli* will demonstrate that the parameters embedded in the framework can indeed be trained to reproduce experimental data. Subsequently, the ability of the framework to serve as a predictive tool will be highlighted with reference to the SOS response system of *E. coli*. Simulation results will focus on DEMSIM's ability to accurately predict the *de novo* response of the system to externally imposed perturbations. Finally, simulation studies for the *araBAD* system will demonstrate the framework's ability to distinguish between different plausible regulatory mechanisms postulated to explain observed gene expression profiles. Overall, the presented results will highlight the broad applicability of the discrete-event paradigm, on which DEMSIM is based, to gene regulatory systems.

64

On the Futility of Optima in Network Inferences and What Can Be Done About It

Charles (Chip) E. Lawrence* (lawrence@dam.brown.edu)

Brown University, Providence, RI

Inference of metabolic and regulatory networks has become a hot topic over the last few years, and a number of reports of efforts to infer such networks using genome scale data have appeared. One factor often overlooked in these fledgling efforts stems from the large size of network solution spaces, which grow exponentially with the number of nodes in graphical representation of the network. The difficulty of drawing inferences in high dimensional setting is a well-recognized problem in statistical learning theory that is often enunciated as the "curse of dimensionality". Most explanations of this curse don't convey well its connection with inferences and tend to be abstract. Here I'll use a Bayesian framework to directly illustrate how difficulties in network inference grows rapidly with network size, illustrate how the curse of dimensionality can so easily render even guaranteed optimal solutions unlikely, and show what can be done about it. Specifically, we will see that the curse stems from the large number of terms in the normalizing constant in Bayes rule, and that the individual terms define opportunities for limiting the curse's ill effects. I'll use a special class of planar networks, whose structure allows for guaranteed optimal solutions and a guaranteed representative sampling, to show how optimal are often misleading, but nevertheless likely network connections and excluded connections can be strongly inferred based on samples from the solution space. I'll also show how incompleteness or inaccuracy of an underlying model of a network can yield optimal solutions that are not even close to right. Background material on Bayesian inference and its connection with inferences based on optimality methods will be given as a aid for those who may not be fully familiar with statistical inference methods.

* Presenting author

65

Whole Community Proteomics Study of an Acid Mine Drainage Biofilm Reveals Key Roles for "Hypothetical" Proteins in a Natural Microbial Biofilm

Jill Banfield*[1] (jill@eps.berkeley.edu), Rachna J. Ram[1], Gene W. Tyson[1], Eric Allen[1], Nathan VerBerkmoes[2,3], Michael P. Thelen[1], Brett J. Baker[1], Manesh Shah[3], Robert Hettich[3], and Robert C. Blake II[4]

[1]University of California, Berkeley CA; [2]University of Tennessee-Oak Ridge National Laboratory, Oak Ridge, TN; [3]Oak Ridge National Laboratory, Oak Ridge, TN; and [4]Xavier University of Louisiana, New Orleans, LA

We are studying relatively simple, low diversity microbial communities associated with extremely acidic, metal-rich mine drainage system to develop an understanding of adaptation, evolution, and the linkages between microbial activity and environmental geochemistry. In order to assess the genetic potential of an entire natural biofilm sample we partially reconstructed the genomes of five dominant organisms (Tyson et al. 2004). Comparative genomic analyses have revealed the structure of each population and provided insights into the processes that create and remove genome heterogeneity. The genomes contain large blocks dominated by genes that encode hypothetical proteins. These are regions inserted in one species or strain relative to others and are inferred to be of phage origin. Phage insertion appears to be the most rapid process leading to strain heterogeneity, and possibly diversification. Within strain populations, gene order is largely retained, and insertions or loss of single genes are rare events. Bacterial populations are dominated by a single clonal type. Homologous recombination is a key force shaping the genomes of archaeal populations in the system, but is rare between different species.

We have characterized the protein complement of a natural microbial community similar to that studied genomically in order to determine which genes are expressed and functionally important. By combining mass spectrometry-based "shotgun" proteomics with community genomics we confidently identified at least 1700 proteins from five dominant species. Proteins involved in protein refolding and response to oxidative stress were abundant, indicating that damage to biomolecules is a key challenge for survival. We validated more than 400 hypothetical proteins, a small subset of which are encoded within blocks of genes apparently acquired by lateral transfer. Entire operons encoding expressed, novel, lineage-specific proteins may be important for acid, metal and radical tolerance. 26% of the detected *Leptospirillum* group II proteins were hypothetical. An extracellular fraction was dominated by a novel protein shown to be a cytochrome central to iron oxidation and AMD formation. Sequencing of DNA encoding cytochrome regions for which peptides were not recovered revealed two amino acid substitutions. Using the strain variant sequence, 100% peptide coverage of the mature protein was achieved. Thus, an iterative genomic and proteomic approach analyses enable detailed in situ analyses of activity within natural microbial consortia.

* Presenting author

66

Application of High Throughput Microcapsule Culturing to Develop a Novel Genomics Technology Platform

Martin Keller[1]* (mkeller@diversa.com), Karsten Zengler[1], Marion Walcher[1], Carl Abulencia[1], Denise Wyborski[1], Sherman Chang[1], Imke Haller[1], Trevin Holland[1], Fred Brockman[2], Cheryl Kuske[3], and Susan Barns[3]

[1]Diversa Corporation, San Diego, CA; [2]Pacific Northwest National Laboratory, Richland, WA; and [3]Los Alamos National Laboratory, Los Alamos, NM

Project Description

The overall goal of this project is to demonstrate the combination of high-throughput cultivation in microcapsules, which gives access to previously uncultivated microorganisms, with genome sequencing from one to a few microcolony-containing microcapsules. This will allow direct access to physiological and genomic information from uncultured and/or difficult-to-culture microorganisms. This approach is fundamentally different than characterization and/or assembly of shotgun or BAC clones derived from community DNA or RNA. The units of analysis in our approach are living, pure microbial cultures in microcapsules, as opposed to the disassembled mixture of small fragments of genomes and cellular networks that have lost their biological context in studies using community nucleic acids. It is envisioned that the microcapsule based, high-throughput cultivation method will also be combined with Transcriptomics and Proteomics technology in the future.

Project achievements

A high-throughput cultivation method based on single cell encapsulation in microcapsules in combination with flow cytometry has been applied to two soil samples. It has been demonstrated that microorganisms belonging to a variety of bacterial phyla, such as *Acidobacteria*, *Gemmatimonadetes* and candidate division TM7, were growing in the microcapsules and subsequently formed microcolonies.

A fluorescent in situ hybridization (FISH) method has been optimized to selectively target and sort encapsulated microcolonies of interest using fluorescence activated cell sorting.

A whole-genome amplification technique has been employed to acquire a sufficient mass of DNA from targeted, encapsulated microcolonies (*E. coli*) to generate libraries for shotgun sequencing of entire genomes. The whole-genome amplification has been optimized so that DNA from as few as two cells can be amplified routinely. The genome coverage of amplified FISH-targeted microcolonies was evaluated using microarrays. Microarray data demonstrated a genome-coverage of 97% to 99% without high levels of bias towards certain genes. Subsequently, genomic libraries have been constructed from these amplified DNA samples. Almost 300 clones of each library have been sequenced. Sequence analysis confirmed that 70% of the clones originated from *E. coli* DNA. Additional gene sequences were affiliated with certain beta-*Proteobacteria* typically found as experimental contaminants.

This work was supported by the Office of Science (BER), U.S. Department of Energy, Grant No. DE-FG02-04ER63771.

* Presenting author

67

Environmental Bacterial Diversity from Communities to Genomes

Janelle R. Thompson[1,2]*, Silvia G. Acinas[1], Vanja Klepac-Ceraj[1,2], Sarah Pacocha[1,2], Chanathip Pharino[1], Dana E. Hunt[1], Luisa A. Marcelino[1], Jennifer Benoit[1,2], Ramahi Sarma-Rupavtarm[1], Daniel L. Distel[3], and Martin F. Polz[1] (mpolz@mit.edu)

[1]Massachusetts Institute of Technology, Cambridge, MA; [2]Woods Hole Oceanographic Institution, Woods Hole, MA; and [3]New England Biolabs, Beverly, MA

We are studying the patterns of diversity among co-occurring coastal bacterioplankton from the level of the entire community to the individual genome. Our goal is to advance the understanding of structure-function relationships in microbial assemblages addressing questions including: What is the range of genomic diversity encompassed by functionally similar populations in specific environmental contexts? What mechanisms govern selection and diversification of natural microbial populations? Using environmental 16S ribosomal RNA gene sequences (ribotypes) as a proxy for bacteria, we have shown that despite a high diversity in the environment, the majority of organisms fall into closely related clusters (<1% 16S rRNA divergence) (1). Such microdiverse sequence clusters are hypothesized to represent functionally-differentiated populations, which arise by selective sweeps (2) and persist because competitive mechanisms are too weak to purge diversity from within them (1).

To examine this hypothesis we quantitatively estimated the genomic diversity within one 16S rRNA microdiversity cluster (*Vibrio splendidus)*. Quantitative PCR analysis (3) over an annual cycle indicated that *V. splendidus* was consistently present as a member of the coastal bacterioplankton community. Vibrio strains were isolated from representative months and the majority were identified as *V. splendidus*. Determination of sequence diversity of a universally distributed protein-coding gene (Hsp60) among all *Vibrio* isolates showed high heterogeneity but confirmed the monophyly of the *V. splendidus* strains. Still greater heterogeneity was revealed when the number of unique genotypes among strains was assayed by pulse field gel electrophoresis (PFGE), moreover, the PFGE analysis provided evidence that a large proportion of genotypes are differentiated by insertions and deletions of large genome fragments. In a set of 12 *V. splendidus* strains genome sizes ranged from 4.5 to 5.6 Mb with only weak correlation of genome size difference to Hsp60 sequence divergence (R = 0.37).

The high degree of heterogeneity among the *V. splendidus* genomes suggest that the average environmental concentration of individual genotypes is astoundingly small. To illustrate this, we divided the QPCR-based estimates of population size of *V. splendidus* in samples taken in Aug 03, Sept 03, and Oct 03 (1,890, 600, and 640 cells/ml, respectively) by the Chao-1 statistical estimates (4) for the number of Hsp60 alleles (125, 94 and 279) and genotypes (465, 553 and 901) in those same samples. The result suggests that unique Hsp60 alleles occurred in the monthly samples at average concentrations of 2 to 15 cells per ml (or at a frequency of 0.3 to 1%) while unique genotypes were present at ~10-fold lower frequency (average concentration for all samples estimated at <1 cell per ml).

The observed pattern of co-existing diversity suggests that purging of genotypes from the population is rare compared to processes introducing variation and that therefore variation persists because it is either favored (e.g., by balancing selection or resource specialization) or neutral. We present ecological considerations to suggest much of the observed diversity may in fact be neutral in an environmental context. Such observations of extreme genomic heterogeneity among closely related individuals have significant implications for the assembly of genome sequences from environmental samples. In addition, if similar patterns of diversity are common to other bacterial populations caution should be exercised in interpreting the extent to which gene complements or even metabolic traits of individual

* Presenting author

isolates may reflect the overall properties of populations. Indeed our results suggest that not only the gene content, but also quantitative abundance and dynamics of individual traits should be considered when evaluating the ecological significance of differences among coexisting genotypes.

References:

1. S. G. Acinas *et al.*, *Nature* **430**, 551-554 (2004).
2. F. M. Cohan, *Annu. Rev. Microbiol.* **56**, 457–487 (2002)
3. J. R. Thompson *et al.*, *Appl. Environ. Microbiol.* **70**, 4103-4110 (2004).
4. J. B. Hughes, J. J. Hellmann, T. H. Ricketts, B. J. M. Bohannan, *Appl. Environ. Microbiol.* **67**, 4399-4406 (2001).

68

Distribution and Variation of *Prochlorococcus* Genotypes Across Multiple Oceanic Habitats

Adam C. Martiny* (martiny@mit.edu), P. K. Amos Tai, Anne W. Thompson, and Sallie W. Chisholm

Massachusetts Institute of Technology, Cambridge, MA

The cyanobacterium *Prochlorococcus* is very abundant in oligotrophic regions of the world's oceans, constituting up to 50% of the cells in the euphotic zone. Thirty-two cultures have been isolated and based on these, a phylogenetic tree has been constructed showing the presence of 6 clades, four low light adapted and two high light adapted. That opens up two questions: (i) To what extent do these cultures represent the phylogenetic space of *Prochlorococcus*? (ii) Are the patterns of genotypic diversity within a given clade similar in field samples collected from different geographical locations?

We are using the sequence of the intergenic transcribed spacer region between the small and large subunit rRNA (ITS) as a neutral marker for genetic variation to describe the diversity of *Prochlorococcus* in field samples. We have cloned and sequenced 1200 ITS fragments from the North Pacific subtropical gyre (Hawaii Ocean Time Series), Sargasso Sea (Bermuda-Atlantic Time Series) at three depths – 25m, 80m and 160m and an upwelling region off the coast of Mexico at two depths (60m and 130m). The phylogenetic analysis showed a high frequency of sequences belonging to the 9312-clade from the 25m samples and 80m — consistent with the finding that the 9312 'ecotype' numerically dominates the upper euphotic zone in many oceanic environments (Zinser et al, in prep.). A comparison using Mantel test of the microdiversity within the 9312 clade at 25m and 80m revealed that the populations were significantly different between these two depths. In addition, a significant amount of "yet to be cultured" diversity was discovered at both 80 and 160m depth including new lineages as well as sub-lineages within the 6 known clades. In a sample collected from a sub-oxic zone off the coast of Mexico, we found a group affiliated to the low light adapted 9313 ecotype, but forming an independent lineage not yet seen in other samples. A future goal is to target such new lineages, amplify their genome using the approach described by Zhang et al. (this meeting) and thereby expand our view on the *Prochlorococcus* physiology and evolution through comparative genome analysis..

* Presenting author

69

From Perturbation Analysis to the Genomic Regulatory Code: the Sea Urchin Endomesoderm GRN

Paola Oliveri*,[1] (poliveri@caltech.edu), Pei-Yun Lee[1], Takuya Minokawa[2], Joel Smith[1], Qiang Tu[1], Meredith Howard[1], David McClay[3], and Eric H. Davidson[1]

[1] California Institute of Technology, Pasadena, CA; [2] Tohoku University, Asamushi, Aomori, Japan; [3]Duke University, Durham, NC

The sea urchin endomesoderm gene regulatory network (GRN) is the most comprehensively understood regulatory apparatus for control of spatial and temporal gene expression in any complex developmental system. It contains almost 50 genes, mainly encoding regulatory proteins. It was initially constructed on the basis of spatial and temporal gene expression data, interpreted through a large scale, systematic, perturbation analysis in which expression of each gene was taken out or otherwise altered, and the effects on all other relevant genes measured quantitatively and with high sensitivity. The GRN provides the essential, overall, "transformation function" by which can be solved the causal relations between the genomic regulatory code that is hardwired into the DNA sequence, and the observed events of spatial and temporal gene expression. That is, it specifies the key *cis*-regulatory inputs into regulatory genes, and their key outputs terminating at other regulatory genes. Hence it is directly testable at the *cis*-regulatory level. In the last year, we identified by computational and experimental methods, and isolated, over a dozen of the central *cis*-regulatory nodes of the GRN. We then require that in gene transfer experiments that these genomic fragments display the same responses to the appropriate perturbations as do the endogenous genes in the whole embryo; and that when the genomic target sites for the relevant inputs are mutated, that the *cis*-regulatory constructs behave in the expected ways. This system wide task will be completed this year, and to date the results indicate that the perturbation analysis indicated the real encoded linkages with perhaps surprising accuracy. In addition: we have demonstrated for the first time that knowledge of the GRN can be used to reengineer the process of development; we have found ways to identify the modular "kernels" of the GRN, which consist of multiple genes recursively "wired" to one another and which are evolutionarily resistant to change; we have developed a new theory of logic processing within *cis*-regulatory modules, as a start on formalization of the genomic regulatory code; we have enlarged our knowledge of the GRN and are adding into it all regulatory genes encoded in the genome that are expressed in the appropriate time and place, so that it will approximate a complete regulatory treatment; and we have begun to extend GRN analysis to later processes.

* Presenting author

Microbial Genomics

70

The Genome of the Ammonia Oxidizing Bacterium *Nitrosomonas europaea*: Iron Metabolism and Barriers to Heterotrophy

Xueming Wei, Neeraja Vajrala, Norman Hommes, Luis Sayavedra-Soto*, and Daniel Arp (arpd@science.oregonstate.edu)

Oregon State University, Corvallis, OR

Nitrosomonas europaea is an aerobic lithoautotrophic bacterium that uses ammonia (NH_3) as its energy source (3). As a nitrifier, it is an important participant in the N cycle, which can also influence the C cycle. The genome sequence of *N. europaea* has been annotated and consists of approximately 2460 protein-encoding genes (1). We are continuing to use the genome sequence to explore the genetic structure and mechanisms underlying the lithoautotrophic growth style of *N. europaea*. Currently, we are investigating its Fe requirements and its possible barriers to utilizing carbon sources different from CO_2.

Because *N. europaea* has a relatively high content of hemes, sufficient Fe must be available in the medium for it to grow. The genome revealed that approximately 5% of the coding genes in *N. europaea* are dedicated to Fe transport and assimilation. Nonetheless, with the exception of citrate, *N. europaea* lacks genes for siderophore production (1). We have initiated the study on this intriguing facet by determining the Fe requirements for growth and are characterizing the expression of the putative membrane siderophore receptors.

N. europaea changes its heme composition when Fe is at a relatively low concentration. Biochemical analyses showed that cytochrome and heme contents of cells grown in Fe-limited medium were 4 fold lower than those from Fe-rich medium. Cellular Fe contents (in both membrane and soluble fractions) showed the same trend. The activity of hydroxylamine oxidoreductase was over three fold lower in cells grown in Fe-limited medium than that in full medium. The growth yields at 0.1 μM Fe and at 0.2 μM Fe were about 35% and 65% respectively of that observed at 10 μM Fe (full medium). *N. europaea* has the mechanisms to cope and grow under Fe limitation.

The *N. europaea* genome revealed that there are over 26 sets of genes that are organized similarly to the genes in a *fecR/fecI* system. Through similarity searches, we have identified possible TonB-dependent receptor genes up- or downstream of these sets. Some of these are similar to genes encoding the siderophore receptors for desferrioxamine (desferal), ferrichrome, and coprogen.

The addition of desferal in Fe-limiting medium promoted the growth of *N. europaea*, though with a longer lag phase, suggesting a necessary induction period of the corresponding receptor. A gene for the putative desferal outer membrane receptor was identified by similarity searches (NE1097, a *foxA* homologue). NE1097 was expressed at a higher level (>10 fold) in Fe-limiting, desferal-containing cultures than in Fe-sufficient cultures. The expression of NE1097 required the presence of desferal, since typical lag phases were observed when inoculants from desferal cultures were used. Several membrane proteins detected only in the cells grown in Fe-limited medium may be involved in Fe

* Presenting author

transport. For example, a membrane peptide with the calculated MW of the putative desferal receptor was observed only in the cells grown in desferal-containing medium. Ferric citrate had an effect similar to that of desferal on *N. europaea* growth in Fe-limiting medium, but with a longer lag phase and a higher final cell density than that in the full medium. Ferrichrome, on the other hand, did not prolong the lag phase, yet increased total cell growth, suggesting that the genes for the ferrichrome receptors were expressed constitutively.

Consistent with the genome sequence data, no siderophores were detected in *N. europaea* culture filtrates under either Fe-limiting or Fe-sufficient conditions using a standard siderophore assay. We considered the possibility that citrate serves as a Fe chelator/siderophore, since *N. europaea* has the necessary genes to produce it. Citrate was detected (2 to 5 μM) in cell-free filtrates from both, low- and full Fe cultures. Surprisingly, cell-free filtrates from full Fe cultures had relatively higher concentrations (5 μM) of citrate than in low Fe cultures (2 to 3 μM). The role of citrate in Fe acquisition, if any, is yet to be determined. *N. europaea* apparently expresses siderophore receptors (i.e. NE1097) under low Fe conditions to scavenge Fe more efficiently. These results reinforce the notion that *N. europaea* uses siderophores produced by other organisms in natural habitats.

Genes encoding the putative outer membrane desferal receptor (NE1097 and NE1088, *foxA* homologues) have been cloned, insertional mutant constructs made, and mutant strains obtained through homologous recombination. Physiological and genetic characterization of these mutants is in progress.

In addition to the Fe experiments, analysis of the *N. europaea* genome has led to experiments probing the possible barriers to heterotrophy in *N. europaea*. The genome contains genes that are similar to the genes encoding fructose transport systems (PTS-type) in other bacteria. Furthermore, *N. europaea* can use fructose as the only source of carbon for growth (2). However, not all the genes required for an active PTS system are present in the genome. The inactivation of the two identified PTS genes did not affect growth on fructose or cause any other growth phenotype. Fructose may enter the cells by some other means. The role of the existing PTS genes remains unclear.

Historically, the activity of the enzyme 2-oxoglutarate dehydrogenase has not been detected in *N. europaea*. The lack of this activity was believed to be the cause for the obligate autotrophy of *N. europaea*. However, the genomic sequence reveals that the three genes necessary to encode this enzyme are present. We inactivated the first gene (odhA) in the operon of this enzyme. The mutant strain grew similarly to wild-type cells during exponential growth. However, in late stationary phase or under ammonia starvation (i.e. energy-limiting conditions), mutant strains lost viability faster and recovered more slowly upon addition of more ammonia as compared to the wild type. This suggests that 2-oxoglutarate dehydrogenase may be involved in processes occurring during the stationary growth phase of *N. europaea*.

A gene encoding a putative ammonia transporter (*amt*) is present in the genome. However, the strain with this gene inactivated showed no difference in growth to wild-type cells over a wide range of ammonium concentrations. The function of *amt* in *N. europaea* is still unknown.

We are exploring the idea that one barrier to heterotrophic growth in *N. europaea* may be due to a lack of transporters for alternative growth substrates. The genes encoding the enzymes to utilize glycerol as the carbon source are present, but the genes encoding a glycerol transporter are not. The heterologous expression of the gene for the glycerol permease from *E. coli* in *N. europaea* permits *N. europaea* to utilize glycerol as the carbon source.

Acknowledgment: This research is funded by grants from the Department of Energy (ER63149-1017305-0007065 and ER63781-1023295-0009824) and resources from the Oregon Agricultural Experiment Station.

* Presenting author

References

1. Chain, P., J. Lamerdin, F. Larimer, W. Regala, V. Lao, M. Land, L. Hauser, A. Hooper, M. Klotz, J. Norton, L. Sayavedra-Soto, D. Arciero, N. Hommes, M. Whittaker, and D. Arp. 2003. Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph *Nitrosomonas europaea*. *J Bacteriol* 185:2759-2773.

2. Hommes, N. G., L. A. Sayavedra-Soto, and D. J. Arp. 2003. Chemolithoorganotrophic growth of *Nitrosomonas europaea* on fructose. *J Bacteriol* 185:6809-6814.

3. Wood, P. M. 1986. Nitrification as a bacterial energy source, p. 39-62. *In* J. I. Prosser (ed.), *Nitrification*. Society for General Microbiology, IRL Press, Oxford.

71

Pelagibacter ubique: A Post-Genomic Investigation of Carbon Metabolism and Photochemistry in an Extraordinarily Abundant Oceanic Bacterium

Stephen J. Giovannoni[1] * (steve.giovannoni@oregonstate.edu), Lisa Bibbs[2], James Tripp[1], Scott Givan[1], Jang-Cheon Cho[1], Martha D. Stapels[3], Russell Desiderio[1], Mercha Podar[2], Kevin L. Vergin[1], Mick Noordeweir[2], Michael S. Rappé[4], Samuel Laney[1], Douglas F. Barofsky[1], and Eric Mathur[2]

[1]Oregon State University, Corvallis, OR; [2] Diversa Corporation, San Diego, CA; [3]Waters Corporation, Milford, MA; and [4]Hawaii Institute of Marine Biology, Kaneohe, HI

The alphaproteobacterium SAR11, now known as *Pelagibacter ubique,* is arguably the most abundant organism in the oceans, where it accounts for approximately 25% of all microbial plankton cells. During summer periods it may exceed 50% of the cells in the surface waters of temperate ocean gyres. *Pelagibacter* plays a key role in the oxidation of the oceanic dissolved organic carbon pool, which is approximately equivalent in size to the atmospheric carbon dioxide pool. The first cultured strains of *Pelagibacter* were isolated by high throughput methods for culturing cells by dilution into natural seawater media, and screening using a new cell array technology. *Pelagibacter* cultures are routinely propagated in autoclaved seawater, where they attain cell densities that are typical of native populations (ca. 10^6 cells/ml). During the sequencing of the 1.3 million base pair genome of *Pelagibacter ubique*, in collaboration with Diversa Corp., it was discovered that this organism has a proteorhodopsin (PR) gene. Liquid chromatography and tandem mass spectrometry were used to prove that the *Pelagibacter* PR gene is expressed in culture and that an identical protein is abundant in coastal Oregon seawater. Laser flash excitation experiments with whole, cultured cells revealed absorption transients with decay kinetics characteristic of retinylidene ion pumps, and light-dependent drops in pH provided confirmation that this PR is a light-dependent proton pump. *Pelagibacter ubique* is the first cultured bacterial isolate to exhibit the PR genes discovered by Bejá, Delong, and coworkers, and the only experimental choice at present for understanding how light-dependent proton pumps influence the efficiency of dissolved organic carbon (DOC) assimilation by heterotrophic bacteria in the ocean surface. The *Pelagibacter* genome is almost exactly the size of the genomes of the obligate intracellular parasites *R. conorii* and *W. pipientis*, but it appears to encode a relatively complete metabolic repertoire governed by unusually simple regulatory circuits. One objective of our current research is to predict the organic carbon sources used by *Pelagibacter* by metabolic reconstruction. Another major thrust of our research is the application of mass spectrometry methods to understand the regulatory responses of *Pelagibacter* to environmental variables, and to explore the proteome state

* Presenting author

of *Pelagibacter* cells in the oceans, so that they can be used as proxies to report the biological state of the system. Metabolic modeling of *Pelagibacter* is an attractive long range goal because it is one of the smallest and simplest cells known. Its remarkable success may be attributable to the integration and optimization of metabolic processes for efficiency at low nutrient fluxes..

72

Does the Three Dimensional Organization of the Nucleoid of the *Deinococcaceae* Contribute to their Ionizing Radiation Resistance?

J. M. Zimmerman and J. R. Battista* (jbattis@lsu.edu)

Louisiana State University and A & M College, Baton Rouge, LA

Transmission electron micrographs of *Deinococcus radiodurans* R1 suggest that the nucleoid of this species exists as a toroidal ring, and have led to speculation that this structure facilitates the extreme radioresistance of this species. However, little direct evidence supports this contention. Since extreme radioresistance is characteristic of all the members of the *Deinococcaceae*, we hypothesize that if nucleoid morphology contributes to radioresistance, the genomic DNA of each species should form similar structures. Using epifluorescence and deconvolution microscopy, we evaluated the nucleoid morphologies of eight of the nine validly described species of *Deinococcus*, the radioresistant bacterium *Rubrobacter radiotolerans*, and the less radioresistant *Thermus aquaticus*, a distant relative of the deinococci. Although the nucleoids of *Deinococcus murrayi, Deinococcus proteolyticus, Deinococcus radiophilus,* and *Deinococcus grandis* have structures similar to *D. radiodurans*, the nucleoids of *Deinococcus radiopugnans* and *Deinococcus geothermalis* lack specific organization. The nucleoid of *R. radiotolerans* consists of multiple highly condensed spheres of DNA. Since only five of the seven recognized deinococcal species exhibit a structurally distinct nucleoid, we conclude there is no obvious relationship between the three dimensional organization of genomic DNA and extreme radioresistance. However, the genomic DNA of all extremely radioresistance species is highly condensed relative to the more radiosensitive species examined. We have examined nucleoid structure following the introduction of DNA double strand breaks and show that the shape of the nucleoid does not demonstrably change in radioresistant species even in strains incapable of repairing strand breaks, suggesting that DNA held in this tightly packed configuration contributes to the radioresistance of these bacteria.

* Presenting author

73

Large Scale Genomic Analysis for Understanding Hydrogen Metabolism in *Chlamydomonas reinhardtii*

Michael Seibert*[1] (mike_seibert@nrel.gov), Arthur R. Grossman[2], Maria L. Ghirardi[1], and Matthew C. Posewitz[3]

[1]National Renewable Energy Laboratory, Golden, CO; [2]Carnegie Institution of Washington, Stanford, CA; and [3]Colorado School of Mines, Golden, CO

Many taxonomically diverse microorganisms have the ability to produce H_2 anaerobically in pathways coupled either to dark fermentation, dark CO-oxidation, light-dependent N_2-fixation, or photosynthetic water oxidation. However, only certain photosynthetic organisms are able to directly couple water oxidation to the photoproduction of H_2. Among these *Chlamydomonas reinhardtii*, a green alga, is able to convert the low potential reductant generated from water by photosynthesis into H_2. A more fundamental understanding of the physiology, biochemistry and genetics of this prototype alga might enable the future development of a sustainable system for H_2 production. Recently, we and our collaborators announced the discovery of a physiological switch (sulfur deprivation), which attenuates photosystem II O_2-evolution activity (Wykoff et al., 1998) and allows *C. reinhardtii* to metabolize O_2 in sealed culture vessels (Melis et al., 2000). This produces an anaerobic environment and leads to the photoproduction of volumetric amounts of H_2 over a 4-day period in batch culture. Under sulfur-deprived conditions, algal cultures are subjected to a mixed metabolic state in which anaerobic fermentation, oxygenic photosynthesis and aerobic respiration co-occur. This physiological state provides us with a unique opportunity to further explore the gene-expression patterns and protein networks that sustain algal H_2 production.

With the completion of the *C. reinhardtii* genome sequence as a part of the DOE Office of Science's Genomics:GTL Program, it is now possible to thoroughly explore large-scale transcript profiles associated with H_2 metabolism in this alga using gene microarrays. High-density DNA microarrays are being used to examine the ways in which WT and mutant strains of *C. reinhardtii* acclimate to conditions that allow for H_2 production. Recently, an array with approximately 3,000 elements was used to examine sulfur deprivation responses in WT and mutant *C. reinhardtii* strains (Zhang et al., 2004). A new array based on specific synthetic 70 mers that represents approximately 10,000 genes has been developed at the Carnegie Institution (Stephan Eberhard and Arthur Grossman, unpublished); it should be ready for use by the beginning of 2005.

We have isolated several independent *C. reinhardtii* mutants with attenuated H_2-photoproduction activity at NREL using a rapid screening technique and will compare gene expression profiles from these mutants with the WT under conditions that facilitate H_2 production. One mutant lacks a functional *HydEF* gene (Posewitz et al., 2004), which encodes a radical SAM protein that is required to insert the metal catalytic center into the hydrogenase enzyme. The *hydEF-1* mutant is the only reported *C. reinhardtii* strain that is unable to produce any H_2 at all. The *hydEF* mutation specifically disrupts hydrogenase activity, but this mutant has no discernable phenotype in comparison to the WT when grown aerobically in the light. Nevertheless, the hydrogenase structural genes are induced during anaerobiosis. These data indicate that the necessary transcriptional, regulatory and signaling pathways required for hydrogenase induction remain intact in this mutant. Consequently, the genes differentially expressed in this mutant relative to the WT, should be a consequence of the mutant's inability to photoproduce H_2.

* Presenting author

Another *C. reinhardtii* mutant, *sta7-10* (Posewitz et al., 2004), is unable to accumulate intracellular starch. Interestingly, this mutant shows aberrant induction of hydrogenase-gene transcription and attenuated H_2-photoproduction activity during anaerobiosis, which correlates with the redox state of its plastoquinone (PQ) pool. One reasonable hypothesis is that the redox state of the PQ pool may signal regulatory processes responsible for turning on or off specific genes involved in anaerobic fermentative pathways and H_2 production.

Previous physiological studies have linked photosynthetic electron transport and fermentation to H_2 production in *C. reinhardtii*. However, a more general knowledge of the metabolic and regulatory context that facilitates H_2 production will be necessary to understand current limitations in H_2-production yields. We will begin to develop a global understanding of the factors that promote H_2 production during anaerobiosis by analyzing transcript profiles from WT and mutant cultures of *C. reinhardtii*. We will also investigate whether intracellular energy stores and/or redox carriers modulate this activity and/or influence the expression of genes needed for H_2 generation. This work will more fully elucidate the biochemical pathways utilized by *C. reinhardtii* during anaerobiosis and provide insights into how mutants altered in normal fermentative metabolism acclimate to anaerobiosis. Genome-wide expression data will also facilitate the modeling of carbon and reductant fluxes, guiding future molecular and metabolic engineering approaches to improve H_2 output by *C. reinhardtii*.

References

1. Melis, A., Zhang, L., Forestier, M., Ghirardi, M. L. and Seibert, M. (2000). Sustained photobiological hydrogen gas production upon reversible inactivation of oxygen evolution in the green alga *Chlamydomonas reinhardtii*. Plant Physiol **122**, 127-36.

2. Posewitz, M. C., King, P. W., Smolinski, S. L., Zhang, L., Seibert, M. and Ghirardi, M. L. (2004). Discovery of two novel radical S-adenosylmethionine proteins required for the assembly of an active [Fe] hydrogenase. J Biol Chem **279**, 25711-20.

3. Posewitz, M. C., Smolinski, S. L., Kanakagiri, S., Melis, A., Seibert, M. and Ghirardi, M. L. (2004). Hydrogen Photoproduction Is Attenuated by Disruption of an Isoamylase Gene in *Chlamydomonas reinhardtii*. Plant Cell **16**, 2151-63.

4. Wykoff, D. D., Davies, J. P., Melis, A. and Grossman, A. R. (1998). The regulation of photosynthetic electron transport during nutrient deprivation in *Chlamydomonas reinhardtii*. Plant Physiol **117**, 129-39.

5. Zhang, Z., Shrager, J., Jain, M., Chang, C.W., Vallon, O. and Grossman, A. R. (2004). Insights into the survival of *Chlamydomonas reinhardtii* during sulfur starvation based on microarray analysis of gene expression. Eukaryot Cell.**3**, 1331-48.

* Presenting author

74

Exploring the Genome and Proteome of *Desulfitobacterium hafniense* DCB2 for its Protein Complexes Involved in Metal Reduction and Dechlorination

James M. Tiedje[1]*, Sang-Hoon Kim[1], Christina Harzman[1], John Davis[2], Brett Phinney[1], Michael Ngowe[1], Washington Mutatu[1], William Broderick[1], David DeWitt[1], Joan Broderick[1], and Terence L. Marsh[1] (marsht@msu.edu)

[1]Michigan State University, East Lansing, MI and [2]Columbus State University, Columbus, GA

The strictly anaerobic bacterium *Desulfitobacterium hafniense* DCB-2 grows by pyruvate fermentation and by alternate respiration using a wide range of electron acceptors including sulfur compounds, chlorinated compounds, and oxidized metals. This organism also grows with N_2 as the sole nitrogen source. The metabolic versatility makes this organism useful for studies of the mechanism of dechlorination and metal reduction, and potentially useful in bioremediation. The sequence of the genome has been determined and seven ORFs similar to the *cprA* (reductive dehalogenase or RDase) gene of *D. dehalogenans* and four ORFs similar to *nifH* of nitrogenase have been detected. To identify the RDase genes induced during alternate respiration with 3-chloro-4-hydroxybenzoate (3C4HBA), 3,5-dichlorophenol (DCP), or *ortho*-bromophenol (*o*-BP), Xeotron® microarrays of the genome of DCB-2 were prepared. Competitive hybridization of cDNA prepared from cultures grown by pyruvate fermentation and under three dehalorespiration conditions indicated that three RDase genes (designated as MENN, MFRS, and MSGV) were induced by 3C4HBA, and two genes (MSSA and VKMN) were induced by both DCP and *o*-BP. Also induced were genes within putative RDase operons, transporter/permease genes, and genes involved in electron transport systems. RT-PCR analysis targeting the seven *cprA* homologs revealed the same patterns of RDase gene expression. Northern hybridization assays targeting three RDase genes (MENN, MSSA, and VKMN) with RNAs from the four culture conditions showed a single mRNA species transcribed from MENN (1.8 kb) and VKMN (2 kb) that is long enough to contain genetic information for two linked genes, the RDase (*cprA*) and the adjacent docking protein (*cprB*) genes. However, Northern hybridization was not adequate as confirmation of the microarray results due to the cross-hybridization of probes to mRNAs from the multiple RDase genes that are phylogenetically related to varying degrees.

The physiology of heavy metal reduction under conditions where the oxidized metal was the only available electron acceptor was investigated with *D. hafniense* DCB-2 growing on a defined minimal freshwater media. The ability of DCB-2 to reduce Fe(III), Cu(II), U(VI), and Se(VI) were each tested separately. Bacterial growth under these metalorespiration conditions were observed for Fe(III), Cu(II), and U(VI), but not Se(VI). Reduction of Se(VI) did occur by DCB-2 when grown by pyruvate fermentation. SEM of DCB-2 morphology under fermentative growth with Se(VI) revealed that selenium is concentrated in polyps attached to the outside of the cell. The oxidation state of this selenium is unknown. Biofilm formation was observed for DCB-2 under conditions of fermentation (DCB-1 media) and respiration (ferric citrate media) when grown on either one of two different beads (Dupont and Siran™). Preliminary morphological examination under light microscope of planktonic cells grown in different heavy metal reduction conditions revealed considerable diversity in cellular morphologies, depending on the specific growth factors. This suggests a diverse adaptive repertoire to varying environmental conditions.

* Presenting author

75

An Integrative Approach to Energy, Carbon, and Redox Metabolism in the Cyanobacterium *Synechocystis* sp. PCC 6803

Wim Vermaas[1,*] (wim@asu.edu), Robert Roberson[1], Allison van de Meene[1], Bing Wang[1], Sawsan Hamad[1], Zhi Cai[1], Julian Whitelegge[2], Kym Faull[2], Sveta Gerdes[3], Andrei Osterman[3], and Ross Overbeek[3]

[1]Arizona State University, Tempe, AZ; [2]University of California, Los Angeles, CA; and [3]Fellowship for the Integration of Genomes (FIG), Argonne, IL

The goal of this project is to merge knowledge from genomic, bioinformatic, proteomic, metabolic, ultrastructural and other perspectives to understand how cyanobacteria live, adapt and are regulated. This project focuses on the cyanobacterium *Synechocystis* sp. PCC 6803, which is spontaneously transformable and has a known genome sequence. Cyanobacteria contribute greatly to global photosynthetic CO_2 fixation and are related to the ancestors of chloroplasts.

Electron tomography and structure determinations. Cyanobacteria have a comprehensive internal membrane system, the thylakoids, where photosynthetic reactions take place. Using electron tomography, we have determined the 3-D structure of cyanobacterial cells with about 4 nm resolution in all directions (also see http://lsweb.la.asu.edu/synechocystis/). Also, we have been successful with high-resolution freeze fracture demonstrating the layered character of the thylakoid membranes and the presence of particles on the outer, cytoplasmic and thylakoid membranes that represent integral and peripheral membrane proteins. The glycocalyx of the cell wall was also exposed. Together these images further expand our 3-D insights in the structure of the cyanobacterial cell and the biogenesis of its components.

Studies on mutants that cannot synthesize chlorophyll in darkness show that such mutants do not have thylakoids in darkness, but will develop thylakoids within hours of illumination. Before illumination, cellular structures are observed that may be thylakoid precursors. In addition, the content of the storage compound polyhydroxyalkanoate (PHA) was determined in wild type and different mutants from 70 nm sections. Initial results were compared with the biochemical data obtained from GC/MS. Particularly in mutants that cannot respire due to the lack of terminal oxidases, polyhydroxybutyrate accumulates, suggesting that this bioplastic serves as a storable fermentation product of this organism. High-resolution, 3-D structural investigations have also led to the discovery of new structures in cyanobacteria. For example, a fine, reticulate network of filaments in the cytoplasm was observed. The nature of these filaments is being determined.

Metabolic fluxes. Another important aspect that is investigated is the physiology of the organism in terms of its carbon fixation and utilization. The relative fluxes of the various central carbohydrate utilization pathways (glycolysis and the pentose phosphate pathway) are investigated by means of ^{13}C-labeling in wild type and mutants impaired in either glycolysis or the pentose phosphate pathway, and grown under various conditions. ^{13}C-glucose is rapidly taken up and converted. Upon isolation of metabolites at different times after ^{13}C-glucose addition, labeled and unlabeled metabolite products are separated by LC and analyzed by MS. The resulting patterns are then used to determine metabolic fluxes through central carbon metabolism pathways. Labeling processes occur on the timescale of 5-40 min, with different compounds showing different labeling kinetics, demonstrating the usefulness of this approach to follow the path and rate of carbon metabolism in vivo.

* Presenting author

Proteome studies. Relative expression and protein turnover studies traditionally employ complete substitution of stable isotopes. This approach has limitations, and we are investigating the use of labeling with 2-4% ^{13}C to code samples for expression proteomics and turnover measurements. Altering the ^{13}C abundance to ~2% yields a measurable effect on the peptide isotopic distribution and the inferred isotope ratio. Elevation of ^{13}C abundance to 4% leads to extension of isotopic distribution and background peaks across every unit of the mass range.

Subtle modification of the isotope ratio (~1-2% increase in ^{13}C) had no effect upon either the ability of data-dependent acquisition software or database searching software to trigger tandem mass spectrometry or match MSMS data to peptide sequences, respectively. More severe modification of the isotope ratio caused a significant drop in performance of both functionalities. Software for deconvolution of isotope ratio concomitant with protein identification using LC-MSMS has been developed (Isosolv). Subtle modification of isotope ratio proteomics (SMIRP) offers a convenient approach to *in vivo* isotope coding.

Bioinformatics. The work of the FIG research team resulted in the deployment of a new open source genomic platform, the SEED. The SEED represents a new generation of software for genome comparative analysis containing one of the largest (and permanently growing) genome collections. The complete system with annotations and tools is freely available. The SEED supports: (1) semi-automated genome comparative analysis and annotation, (2) pathway and subsystem reconstruction and analysis across multiple species, (3) community annotations and alternative assignments from major public integrations, (4) gene discovery using genome context analysis techniques, and (5) integration, comparative analysis and interpretation of functional genomics data. The current integration (http://theseed.uchicago.edu/FIG/index.cgi) contains data from 470 bacterial (of these 261 are complete or near completion), 32 archaeal (21 more or less complete), 558 (16 complete) eukaryotic, and 1272 viral genomes, including complete and nearly complete genomes of 14 cyanobacteria, as well as anoxygenic phototrophic bacteria and higher plants – invaluable for comparative genomic studies of energy and carbon metabolism in *Synechocystis* sp. PCC 6803. An important unique feature of the SEED is the support of metabolic reconstruction and comparative genome analysis via encoding and projection of functional subsystems. The FIG research team has validated the new software by developing over 150 core subsystems, covering many aspects of central metabolism. Another important SEED feature is that it is readily editable and expandable by an inexperienced user. The editing of existing subsystems and the construction of new ones is straightforward and does not require any programming skills. This provides experimental biologists with unique opportunities of fully interactive *in silico* analysis of metabolic pathways at a whole-genome scale and distinguishes the SEED from other valuable resources, such as KEGG.

The SEED platform provided the foundation for the development of CyanoSEED (to be released in January 2005), a specialized portal to comparative analysis, community-based annotation, and metabolic reconstruction of all available cyanobacterial genomes. Many new subsystems covering areas specific for cyanobacteria were added to the CyanoSEED, including: cyanobacterial photosynthetic and respiratory membrane complexes, inorganic carbon concentration and fixation mechanisms, several pigment and cofactor biosynthetic pathways, etc. A comprehensive metabolic reconstruction effort supported by the CyanoSEED provides all components required for compiling stoichiometric matrices and starting flux-balance modeling.

This combination of molecular and cell biology, genomics, proteomics and metabolome analysis leads to comprehensive insight in cyanobacterial physiology and structure, and helps to elucidate the workings of this ecologically and evolutionarily important group of organisms.

* Presenting author

76

Role of Cellulose Binding Modules in Cellulose Hydrolysis

David B. Wilson* (dbw3@cornell.edu) and Shaolin Chen

Cornell University, Ithaca, NY

Cellulose binding modules (CBM) are found on most cellulases that catalyze the hydrolysis of crystalline cellulose, as well as on many hemicellulases produced by cellulolytic microorganisms. Removal of a CBM usually does not reduce activity on soluble substrates, such as CMC or cellulodextrins, but significantly reduces activity on crystalline cellulose (1). There is clear evidence that a major role of a CBM is to keep the catalytic domain close to its substrate, thus increasing its ability to bind to and hydrolyze individual cellulose chains (2). However, there also have been several reports that CBMs can disrupt the surface of cellulose, presumably by breaking some of the hydrogen bonds, which hold chains together (3,4). However, other workers have not seen this and thus this second role is still controversial (2, 5).

We had shown that addition of a hundred fold molar excess of a family 2 CBM to the *Thermobifida fusca* endocellulase Cel6A catalytic domain (lacking its CBM) did not give any stimulation of its activity on filter paper and thus I did not think that CBMs could disrupt cellulose (5). However, in recent experiments, we found that while the activity of Cel6Acd on filter paper is not stimulated by free CBM, the activity of native Cel6A was doubled by a one hundred-fold molar excess of a family 2 CBM. Furthermore, a twenty-fold excess of *T. fusca* E7 also doubled the activity of Cel6A on filter paper. E7 is a *T. fusca* 18K extracellular protein which is induced by growth on cellulose, is present in large amounts in the culture supernatant and that binds well to cellulose and to chitin. It has weak homology to some family 3 CBMs. These results do provide strong evidence that CBMs can alter the structure of cellulose in a way that makes it more easily hydrolyzed by an endocellulase. Experiments are under way to see if other cellulases are stimulated by E7 and if we can detect the nature of the change in cellulose structure caused by E7 binding.

References

1. Boraston, A.B., Bolam, D.N., Gilbert, H.J., and Davies, G.J. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochem. J. 382: 769-781 (2004).
2. Bolam, D.N., Ciruela, A., McQueen-Mason, S., et al. Pseudomonas cellulose-binding domains mediate their effects by increasing enzyme substrate proximity. Biochem. J. 331: 775-781 (1998).
3. Din, N., Gilkes, N.R., Tekant, B., Miller, R.C. Jr., Warren, R.A.J., and Kilburn, D.G. Non–hydrolytic disruption of cellulose fibres by the binding domain of a bacterial cellulase. Bio/Technology 9: 1096-1099 (1991).
4. Levy, I., Shani, Z., and Shoseyov, O. Modification of polysaccharides and plant cell wall by endo-1,4-beta-glucanase and cellulose-binding domains. Biomol. Eng. 19: 17-30 (2002).
5. Barr, B.K. Hydrolysis, specificity and active site binding of Thermobifida fusca cellulases. PhD Thesis Cornell University (1997)

* Presenting author

77

Three *Prochlorococcus* Cyanophage Genomes: Signature Features and Ecological Interpretation

Matthew B. Sullivan[1]* (mbsulli@mit.edu), Maureen Coleman[1], Peter Weigele[1], Forest Rohwer[2], and Sallie W. Chisholm[1]

[1]Massachusetts Institute of Technology, Cambridge, MA and [2]San Diego State University, San Diego, CA

The oceanic cyanobacteria *Prochlorococcus* are globally important, ecologically diverse primary producers. Their viruses (phages) are thought to mediate population sizes and affect the evolutionary trajectories of their hosts. Here we present an analysis of genomes from three *Prochlorococcus* phages — a podovirus and two myoviruses. Although *Prochlorococcus* are evolutionarily and physiologically distinct from the majority of hosts of previously sequenced phages, the morphology, overall genome features and gene content suggest these phages are quite similar to the T7-like (P-SSP7) and T4-like (P-SSM2 and P-SSM4) phages. Using the existing phage taxonomic framework as a guideline, we examined genome sequences to establish 'core' genes for each phage group and found our cyanophages contained 15 of 26 'core' T7-like genes (P-SSP7) and 43 and 42 of 75 'core' T4-like genes (P-SSM2 and P-SSM4). Outside of these 'core' phage genes, taxonomy of best hits analyses suggest each genome contains a significant number of 'cyanobacterial' genes – i.e., genes which are common in cyanobacteria, but have not been observed among phages outside the cyanophages – some of which we speculate represent 'core' cyanophage genes. For example, all three phage genomes contain photosynthetic genes (*psbA, hliP*) that are thought to help maintain host photosynthetic activity during infection, as well as an aldolase family gene (*talC*) that suggests alternative routes of carbon metabolism are important during cyanophage infection. The podovirus P-SSP7 genome also contains an integrase gene (*int*) and genome features that suggest it is capable of integrating into its host. If functional, this would be the first report of a cultured temperate T7-like phage or a temperate marine cyanophage, and would have significant evolutionary and ecological implications for both phage and host. Further, both myoviruses P-SSM2 and P-SSM4 contain phosphate-inducible genes (*phoH* and *pstS*) that may be important for phage and host responses to phosphate stress, a commonly limiting nutrient for marine cyanobacterial growth. Thus, these marine cyanophages appear to be variations of two well-known phages, but contain genes that, if functional, may document how cyanophages have specialized for infection of photosynthetic hosts in low nutrient oceanic environments.

* Presenting author

78

The Alternative Sigma Factor RpoN Regulon of *Rhodopseudomonas palustris*

Yasuhiro Oda[1]* (yasuhiro-oda@uiowa.edu), Sudip K. Samanta[1], Frank W. Larimer[2], and Caroline S. Harwood[1]

[1]University of Iowa, Iowa City, IA and [2]Oak Ridge National Laboratory, Oak Ridge, TN

The alternative RNA polymerase sigma factor RpoN activates a wide range of genes involved in various cellular processes in many different bacterial species. These include nitrogen and carbon metabolism, flagellar biosynthesis, and virulence. The photosynthetic bacterium *Rhodopseudomonas palustris* is an extremely metabolically versatile species. Under anaerobic conditions it can generate energy from light and convert nitrogen gas to ammonia and hydrogen (a biofuel) by nitrogen fixation. It can also degrade lignin monomers. This metabolic versatility is reflected in the genome sequence of *R. palustris* strain CGA009. To address the question of what physiological processes are controlled by RpoN in *R. palustris*, we isolated an *rpoN* Tn5 transposon mutant. The mutant grew in medium containing ammonium as a nitrogen source with the same growth rate as wild-type, but did not grow under nitrogen-fixing conditions. In addition, *rpoN* appears to be involved in motility, hydrogen recycling, and biofilm formation. Introduction of a plasmid containing the *rpoN* gene into the mutant *in trans* complemented all of these phenotypes. To assess the RpoN regulon of *R. palustris* in more detail with an emphasis on nitrogen metabolism, the whole genome gene expression profile of wild-type cells was compared to that of the *rpoN* mutant. The two strains were grown in medium containing yeast extract as the nitrogen source, which derepresses nitrogenases in wild-type cells and allows growth of the *rpoN* mutant. Wild-type cells expressed over 400 genes at levels of 2-fold or higher as compared to the *rpoN* mutant. Among these were the genes involved in nitrogen (e.g., nitrogenases, nitrogen regulatory proteins, ammonium transporters, and glutamine synthetases) and carbon (e.g., lignin monomers, fatty acids, and dicarboxylic acids) metabolism, flagellar biosynthesis, and various transport systems. These results and a computational analysis of the RpoN regulon have elucidated the functional role of the alternative sigma factor RpoN in the successful metabolic opportunist *R. palustris*.

* Presenting author

79

Integrative Control of Key Metabolic Processes in *Rhodopseudomonas palustris* for the Enhancement of Carbon Sequestration and Biohydrogen Production

F. Robert Tabita[1]* (Tabita.1@osu.edu), Janet L. Gibson[1], Caroline S. Harwood[2], Frank Larimer[3], J. Thomas Beatty[4], James C. Liao[5], and Jizhong (Joe) Zhou[3]

[1]Ohio State University, Columbus, OH; [2]University of Iowa, Iowa City, IA; [3]Oak Ridge National Laboratory, Oak Ridge, TN; [4]University of British Columbia, Vancouver, BC; and [5]University of California, Los Angeles, CA

The nonsulfur purple (NSP) photosynthetic (PS) bacteria (1) are the most metabolically versatile organisms found on Earth and they have become model organisms to understand the biology of a number of important life processes. One bacterium, *Rhodopseudomonas palustris*, is unique in that it is able to catalyze more processes in a single cell than any other member of this versatile group. Thus, this organism probably catalyzes more fundamentally and environmentally significant metabolic processes than any known living organism on this planet. *R. palustris* is a common soil and water bacterium that can make its living by converting sunlight to cellular energy and by absorbing atmospheric carbon dioxide and converting it to biomass. It is often the most abundant NSP PS bacterium isolated in enrichments. Its abundance is most probably related to one of its unique characteristics; i.e., unlike other NSP PS bacteria, *R. palustris* can degrade and recycle components of the woody tissues of plants (wood contains the most abundant polymers on earth). *R. palustris* can do this both aerobically in the dark and anaerobically in the light. Recent work has shown that regulation of the processes of CO_2 fixation, N_2 fixation, and H_2 metabolism is linked in NSP bacteria (2). Moreover, a different, yet uncharacterized regulatory mechanism operates under aerobic conditions (unpublished results). Now that its genome sequence is available through the efforts of the JGI and the members of this consortium (3), interactive metabolic regulation of the basic CO_2, hydrogen, nitrogen, aromatic acid, and sulfur pathways of *R. palustris*, as well as other important processes, can be probed at a level of sophistication that was not possible prior to the completion of the genomic sequence. We have pooled the collective expertise of several investigators, using a global approach to ascertain how all these processes are regulated in the cell at any one time. These studies take advantage of the fact that *R. palustris* is phototrophic, can fix nitrogen and evolve copious quantities of hydrogen gas, and is unique in its ability to use such a diversity of substrates for both autotrophic CO_2 fixation (i.e., H_2, H_2S, $S_2O_3^{2-}$, formate) and heterotrophic carbon metabolism (i.e., sugars, dicarboxylic acids, and aromatics, plus many others) under both aerobic and anaerobic conditions.

With regard to the integrative control of metabolism, we have shown that the control of CO_2 fixation is superimposed on the control of nitrogen fixation and hydrogen metabolism in this organism. By interfering with the normal means by which *R. palustris* removes excess reducing equivalents generated from the oxidation of organic carbon, strains were constructed in which much of the electron donor material required for growth was converted to hydrogen gas. The resultant strains were shown to be derepressed for hydrogen evolution such that copious quantities of H_2 gas were produced under conditions where the wild-type would not normally do this. As *R. palustris* and related organisms have long been proposed to be useful for generating large amounts of hydrogen in bio-reactor systems, the advent of these newly isolated strains, in which hydrogen production is not subject to the normal control mechanisms that diminish the wild-type stain, is quite significant. Moreover,

* Presenting author

R. palustris is unique amongst the nonsulfur purple bacteria in that it is capable of degrading lignin monomers and other waste aromatic acids both anaerobically and aerobically. Inasmuch as the degradation of these compounds may be coupled to the generation of hydrogen gas, by combining the properties of the hydrogen- producing derepressed strains, with waste organic carbon degradation, there is much potential to apply these basic molecular manipulations to practical advances. To maximize this capability, the coordinated application of gene expression profiling (transcriptomics), proteomics, carbon flux analysis and bioinformatics approaches have been combined with traditional studies of mutants and physiological/biochemical characterization of cells. During the course of these studies, novel genes and regulators were identified from investigating control of specific processes by conventional molecular biology/biochemical techniques. These studies, along with the microarray studies discussed above, have shown that there are key protein regulators that control many different processes in this organism. In many instances, further surprises relative to the role of known regulators, such as the Reg system and CbbR, were noted in *R. palustris*. A novel phospho-relay system for controlling CO_2 fixation gene expression was also identified and biochemically characterized and the means by which this system influences other aspects of metabolism is also under study. This latter system, where key regulators contain motifs that potentially respond to diverse metabolic and environmental perturbations, suggests an exquisite means for controlling key processes such as CO_2 fixation. Likewise, interesting and important genes and proteins that control sulfur oxidation, nitrogen fixation, hydrogen oxidation, and photochemical energy generation have been identified and characterized, and the biochemistry of these systems is under intense study.

In summary, functional analysis of the *R. palustris* proteome and transcriptome, along with traditional biochemical/physiological characterization, has led to considerable progress, placing our group in excellent position to address long term goals of computational modeling of metabolism such that carbon sequestration and hydrogen evolution might be maximized.

References

1. Tabita, F. R., and Hanson, T. E. Anoxygenic photosynthetic bacteria. 2004. In: Microbial Genomics. C. M. Fraser, K. E. Nelson, and T. D. Read (eds.). Humana Press, Inc., Totowa, NJ, pp. 225-243.
2. Dubbs, J. M., and Tabita, F. R. Regulators of nonsulfur purple phototrophic bacteria and the interactive control of CO_2 assimilation, nitrogen fixation, hydrogen metabolism and energy generation. FEMS Microbiol. Rev. **28** (2004) 353-376.
3. Larimer, F.W., Chain, P., Hauser, L., Lamerdin, J., Malfatti, S., Do, L., Land, M., Pelletier, D.A., Beatty, J.T., Lang, A.S., Tabita, F. R., Gibson, J.L., Hanson, T. E., Bobst, C., Torres y Torres, J., Peres, C., Harrison, F.H., Gibson, J., and Harwood, C.S. Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris*. Nature Biotechnology **22** (2004) 55-61.

* Presenting author

80

Whole Genome Transcriptional Analysis of Toxic Metal Stresses in *Caulobacter crescentus*

Gary L. Andersen*[1] (GLAndersen@lbl.gov), Ping Hu[1], Eoin L. Brodie[1], and Harley H. McAdams[2]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA and [2]Stanford University School of Medicine, Stanford, CA

Potentially hazardous levels of heavy metals have dispersed into subsurface sediment and groundwater in a number of metal contaminated DOE sites and represent a challenge for environmental restoration. Effective bioremediation of these sites requires knowledge of genetic pathways for resistance and biotransformation by component organisms within a microbial community. The aquatic bacterium *Caulobacter crescentus* is a ubiquitous organism with a distinctive ability to survive in low nutrient environments. It has been selected for extensive study by DOE because of its ability to survive in broad environmental habitats where contamination may be present. The recently completed sequence of the strain CB15 has provided the information to study genome wide response to heavy metal stress. A customized 500,000-probe Affymetrix array has been designed by the McAdams laboratory at Stanford University to measure transcription levels of all 3763 putative ORFs, both strands of hypothetical proteins as well as the intervening intergenic regions. This study used this microarray to study transcriptional response to heavy metal stress.

We studied the toxic effect of six heavy metals (seven compound: methylmercury chloride, cadmium sulfate, sodium selenite, lead nitrate, potassium chromate, potassium dichromate and uranyl nitrate) on growth, survival and cell morphology. We unexpectedly found that strain CB15N was not significantly affected for growth at 1 mM uranium concentration. The highest level of uranium currently observed in ground water at the Oakridge FRC is 200 μM. Under the same conditions in our laboratory, growth of *E. coli* K-12 was completely stopped and the growth of *Pseudomonas putida* (*Pseudomonas spp*. has been reported to accumulate uranium) was drastically reduced. We believe this is the first study to identify *C. crescentus* as a uranium-resistant bacterium. Whole genome transcriptional analysis using the Affymetrix *C. crescentus* microarray revealed groups of genes, operons and pathways, which were up regulated under different heavy metal stresses. Some of the up-regulated pathways (such as DNA repair, removal of superoxide radicals, thio-group protection) confirmed what is known about heavy metal stress on other organisms. Nine transcripts were commonly up-regulated when the cells were stressed with four different toxic metals. We also observed the up-regulation of specific regulatory genes as well as genes and operons of unknown function in response to specific metal stresses. In cells stressed with uranium we observed the up-regulation of four proteins that belong to two different two-component signal transduction systems. Their involvement in uranium stress was confirmed in phenotypic studies by deletion mutants of one signaling pathway. We also identified groups of genes and operons of unknown functions, including transcripts from antisense strand of a predicted gene. Further studies may elucidate function of these transcripts and, ultimately, the mechanism used by *C. crescentus* to overcome uranium toxicity. Whole genome transcriptional analysis provided a powerful tool for the detection of candidate genes, with no prior knowledge, that may be involved in metal stress survival. Such analysis will be increasingly necessary as more microbial genome sequences are completed with only computational annotation to suggest function.

* Presenting author

81

Systematic Analysis of Two-Component Signal Transduction Systems Regulating Cell Cycle Progression in *Caulobacter crescentus*

Michael Laub* (Laub@CGR.Harvard.edu)

Harvard University, Cambridge, MA

Progression through the cell cycle requires the precise coordination of DNA replication, chromosome segregation, cell division, and cell growth. How these processes are coordinated and regulated can be studied in the experimentally tractable model bacterium Caulobacter crescentus. Cell cycle progression in Caulobacter is accompanied by a series of morphological transitions which culminate in the production of two asymmetric daughter cells, a sessile stalk cell that immediately initiates a new round of DNA replication after cell division and a motile swarmer cell that must differentiate into a stalked cell before initiating DNA replication. We have begun a systematic analysis of the signaling and regulatory genes controlling the Caulobacter cell cycle, focusing primarily on the two-component signal transduction systems, comprised of histidine kinases and response regulators. Two-component signaling systems are ubiquitous regulatory pathways in prokaryotes that provide a versatile means of detecting and responding to changes in environmental, cellular, and developmental conditions.

We systematically generated deletion strains for each of the 107 two-component signaling genes (63 histidine kinases and 44 response regulators) encoded in the Caulobacter genome. The systematic phenotypic characterization of these mutants has identified four new two-component genes essential for viability and 16 others required for proper cell cycle progression. The deletion mutants generated were individually bar-coded to enable high-throughput analysis of growth and fitness under a variety of environmental conditions; use of this assay to identify the stimuli for specific two-component systems will be presented. Finally, we have developed a technique, termed kinase-substrate profiling, which allows the rapid and accurate delineation of phosphate flow through two-component signaling pathways. This technique has been employed to identify kinase-regulator pairs among the newly identified cell cycle regulatory proteins. Examples will be presented to demonstrate how the combination of these systematic genetic, biochemical, and genomic approaches can quickly lead to the identification of signaling pathways controlling key cellular processes and metabolic changes.

* Presenting author

82

The U.S. DOE Joint Genome Institute Microbial Program

David Bruce[1]* (dbruce@lanl.gov), Alla Lapidus[2], Patrick Chain[3], Jeremy Schmutz[4], Frank Larimer[5], Nikos Kyrpides[2], Paul Gilna[1], Eddy Rubin[2] and Paul Richardson[2]

[1]JGI-Los Alamos National Laboratory, Los Alamos, NM; [2]JGI-Production Genomics Facility and Lawrence Berkeley National Laboratory, Berkeley, CA; [3]JGI-Lawrence Livermore National Laboratory, Livermore, CA; [4]JGI-Stanford Human Genome Center, Palo Alto, CA; and [5]JGI-Oak Ridge National Laboratory, Oak Ridge, TN

The Department of Energy initiated the Microbial Genome Program (MGP, http://microbialgenome.org) in late 1994 as a spin-off of the Human Genome Program. The principle goal of the MGP is to fund research into microbes related to DOE interests. The DOE Joint Genome Institute (JGI, www.jgi.doe.gov) is composed 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, as well as the Stanford Human Genome Center. The JGI's Production Genomics Facility in Walnut Creek, California is a high throughput sequencing center and the principal engine for the JGI. Originally designed to sequence the DOE commitment to the Human Genome Program (Chromosomes 5, 16, and 19), this facility now routinely produces over 2.5 billion raw bases per month from a wide variety of organisms including microbes and microbial communities. The JGI is responsible for sequencing, assembling, and annotating microbial genomes of interest to the DOE through the MGP and GTL programs.

The JGI Microbial Program was recently established to better coordinate and leverage the capabilities of its partner organizations. A work flow procedure has been formalized to process samples from DNA prep through sequencing, assembly, finishing, quality assurance, annotation and analysis. To date, the JGI has sequenced over 100 microbes to draft quality, finished over 30 and is currently working on more than 60 additional microbial projects. Most projects are now targeted for finishing at one of three JGI locations.

Virtually all microbial projects are sequenced by the whole genome shotgun method. The JGI randomly shears the purified DNA under different conditions and selects for three size populations. Fragments are end repaired and selected for inserts in the range of 3kb, 8kb, and 40kb. These are cloned into different vector systems and checked for quality by pcr or sequencing. Once libraries have passed the QC step, a total coverage of approximately 8.5X sequencing is performed on colonies from the three libraries. The resulting reads are trimmed for vector sequences and assembled. This production sequencing assembly is quality checked and then released to the collaborating PI as the initial Quality Draft assembly and is automatically annotated by the group at Oak Ridge National Laboratory. At this point, the draft assembly is then assigned to a finishing group that will close all gaps, resolve repeat discrepancies, and improve low quality regions. The final assembly is then passed to the Stanford Quality Assurance group to assess the integrity and overall quality of the genome sequence. The finished sequence then receives a final annotation and this package is used as the basis for analysis and publication.

* Presenting author

83

Identification of Genes that are Required for Recycling Reducing Power during Photosynthetic Growth

Christine L. Tavano, Angela M. Podevels, and Timothy J. Donohue* (tdonohue@bact.wisc.edu)

University of Wisconsin, Madison, WI

Photosynthetic organisms have the unique ability to conserve the energy in light and generate reducing power. We are studying what is required for photosynthesis in the α-proteobacterium, *Rhodobacter sphaeroides*. Global gene expression analysis has shown that RNA levels from ~50 genes of unknown function were regulated by changes in light intensity and oxygen tension, like those of known components of the *R. sphaeroides* photosynthetic apparatus. Several of these uncharacterized genes were located in the RSP4157-4164 gene cluster on plasmid P004. A mutant containing a polar insertion in RSP4157, CT01, was able to grow via photosynthesis under autotrophic conditions using H_2 as an electron donor and CO_2 as a carbon source. However, CT01 was unable to grow photoheterotrophically in a succinate based medium unless compounds that can be used to recycle reducing power (the external electron acceptor DMSO or CO_2) were present, suggesting that this mutant was defective in recycling reducing power during photosynthetic growth when a fixed carbon source was present. CT01 had decreased levels of RNA for genes encoding putative glycolate degradation functions. Exogenous glycolate also rescued photoheterotrophic growth of CT01, leading us to propose that CO_2 produced from glycolate metabolism can be used by the Calvin cycle to recycle reducing power generated in the photosynthetic apparatus. The ability of glycolate, CO_2, or DMSO to support photoheterotrophic growth of CT01 suggests that products of the RSP4157 gene cluster serve a heretofore unknown role in recycling reducing power under photosynthetic conditions..

84

A Tightly-Regulated Oscillatory Circuit Formed by Conserved Master Regulator Proteins Controls the *Caulobacter* Cell Cycle

Harley McAdams* (hmcadams@stanford.edu) and Lucy Shapiro

Stanford University School of Medicine, Stanford, CA

A complex oscillatory genetic circuit controls *Caulobacter crescentus* cell-cycle progression and asymmetric polar morphogenesis. Two tightly regulated master regulatory proteins, CtrA and GcrA, were recently shown to form the core oscillator.[1] Two switches triggered by the progression of chromosome replication and cytoplasmic compartmentalization, respectively, act to create a type of escapement mechanism that paces and synchronizes progression of the cell cycle.[1,2] The intracellular concentrations of GcrA and CtrA exhibit both temporal and spatial oscillations that act to activate or repress numerous cell cycle regulated genes. Many of these genes are themselves top-level regulators of modular functions that execute the functions involved in cell cycle progression (for example, replicating the chromosome, initiating the FtsZ ring, or constructing polar structures)

* Presenting author

The architecture of the bacterial cell's regulatory control system is a hierarchical, modular, asynchronous, self-timed control system.[3] In this system, synchronization of the ordering of cell cycle events is dependent on cross module dependencies, that is, checkpoints, often involving signaling via two-component systems (see Michael Laub's abstract). Recent results elaborating this circuit include characterization of the regulons of two additional key *Caulobacter* cell cycle regulatory proteins, DnaA and DivK (publications in preparation). We have also identified and characterized two long-sought proteins centrally involved in active regulation of CtrA proteolysis (publication in preparation).

Caulobacter divides asymmetrically, producing daughter cells with differing polar structures, different cell fates, and asymmetric regulation of the initiation of chromosome replication. Complex intracellular signaling is required to keep the organelle developmental processes at the cell poles synchronized with other cell cycle events. Two recently characterized switch mechanisms controlling cell cycle progress are triggered by relatively large scale developmental events in the cell: the progress of the DNA replication fork and the physical compartmentalization of the cell that occurs well before division. These mechanisms invoke rapid, precisely timed, and even spatially differentiated regulatory responses at important points in the cell cycle. Assays of the relative timing of cytoplasmic compartmentalization and cell division coupled with high resolution 3D tomographic images of the end stages of cell division has shown how breaking of phosphor-signaling paths by blocking the cytoplasmic diffusion of phosphorylated signal proteins leads to differential regulatory programming of the daughter cells.[2,4]

1. Holtzendorff, J., Hung, D., Brende, P., Reisenauer, A., Viollier, P. H., McAdams, H. H., Shapiro, L. Oscillating global regulators control the genetic circuit driving a bacterial cell cycle. *Science* **304**, 983-7 (2004).
2. McGrath, P.T., Viollier, P. & McAdams, H.H. Setting the Pace: Mechanisms tying *Caulobacter* cell-cycle progression to macroscopic cellular events. *Current Opinion in Microbiology*, **7(2)**:192-7. (2004).
3. McAdams, H.H. & Shapiro, L. A bacterial cell-cycle regulatory network operating in time and space. *Science* **301**, 1874-7 (2003).
4. Judd, E.M., Ryan, K., Moerner, W.E., Shapiro, L. & McAdams, H.H. Fluorescence bleaching reveals asymmetric compartment formation prior to cell division in *Caulobacter*. *Proc. Natl. Acad. Aci. USA* **100**, 8235-8240 (2003).

85

Dynamics and Control of Biofilms of the Oligotrophic Bacterium *Caulobacter crescentus*

Alfred M. Spormann (spormann@stanford.edu) and Plamena Entcheva-Dimitrov

Stanford University, Stanford, CA

Caulobacter crescentus is an oligotrophic α-proteobacterium with a complex cell cycle involving sessile stalked and piliated, flagellated swarmer cells. Because the natural lifestyle of *C. crescentus* intrinsically involves a surface-associated, sessile state, we investigated the dynamics and control of *C. crescentus* biofilms developing on glass surfaces in a hydrodynamic system. In contrast to biofilms of the well studied *Pseudomonas aeruginosa*, *E. coli*, and *Vibrio cholerae*, *C. crescentus* CB15 cells form biphasic biofilms, consisting predominantly of a cell monolayer biofilm and a biofilm containing densely-packed, mushroom-shaped structures. Based on comparisons between the *C. crescentus* strain CB15 wild type and its holdfast (*hfsA*, ΔCC0095), pili (Δ*pilA-cpaF*::Ω*aac3*), motility (*motA*), flagellum (*flgH*) mutants and a double mutant lacking holdfast and flagellum (*hfsA*; *flgH*), a model for biofilm formation in *C. crescentus* is proposed: For both biofilm forms, the holdfast structure at the tip of a stalked cell is crucial for mediating the initial attachment. Swimming motility by means of the single polar flagellum enhances initial attachment and enables progeny swarmer cells to escape from the monolayer biofilm. The flagellum structure contributes also to maintaining the mushroom structure. Type IV pili enhance but are not absolutely required for the initial adhesion phase. However, pili are essential for forming and maintaining the well-defined three-dimensional mushroom-shaped biofilm. The involvement of pili in mushroom architecture is a novel function for type IV pili in *C. crescentus*. These unique biofilm features demonstrate a spatial diversification of the *C. crescentus* population into a sessile, as 'stem cells' acting subpopulation (monolayer biofilm), which generates progeny cells capable of exploring the aqueous, oligotrophic environment by swimming motility, and a subpopulation accumulating in large mushroom structures.

86

Widespread and Abundant CelM Endoglucanases of Marine *Cytophaga*-like Bacteria Revealed by Whole Genome Shotgun Sequencing and Fosmid Cloning

Matthew T. Cottrell and David L. Kirchman* (kirchman@cms.udel.edu)

University of Delaware, Lewes, DE

Culture-independent approaches for investigating the composition of marine bacterial communities have revealed that the most abundant members of natural bacterial communities differ from those that have been isolated in culture. Data obtained with a variety of approaches, including fluorescence in situ hybridization (FISH) and fosmid cloning of environmental DNA, have identified *Cytophaga*-like as an abundant constituent of marine bacterial consortia. In our work we have been focusing the role of *Cytophaga*-like bacteria in carbon cycling in the ocean and their adaptations for

* Presenting author

the consumption of dissolved organic material (DOM) in the form of high molecular weight polysaccharides. Previous investigations revealed that in natural marine consortia *Cytophaga*-like bacteria are superior competitors for the consumption of radiolabeled protein, chitin and possibly other polysaccharides. In this study we sought to identify what types of endoglucanases are employed by marine *Cytophaga*-like bacteria to utilize high molecular weight polysaccharides. Our approach was to search contigs bearing *Cytophaga*-like 16S rDNA in the recently published whole genome shotgun sequence database of the Sargasso Sea for endoglucanase genes.

We examined 27 contigs bearing *Cytophaga*-like 16S rDNA, including 11 annotated as such, plus 16 additional contigs we identified as having *Cytophaga*-like 16S rDNA by BLAST analysis of a 16S rDNA data base. BLAST analysis of the *Cytophaga*-like sequences revealed two contigs with open reading frames encoding proteins that are most similar to CelM in the *Cytophaga hutchinsonii* genome sequence. Subsequent BLAST analysis of the 1,001,987 conceptual peptides against a database of 113 cellulases (EC 3.2.1.4) obtained from Swiss-Prot revealed 30 contigs encoding CelM-like cellulases, which was the most prevalent cellulase detected. Cellulases belonging to glycosyl hydrolase family 5 were detected on 27 contigs, while 10 or fewer contigs possessed cellulases similar to those in glycosyl hydrolase families 8, 9, 10 and 12.

PCR primers that amplify a 1,132 bp fragment from selected environmental CelM sequences were used to screen a fosmid library of bacterial DNA from the Arctic Ocean. The CelM-positive fosmid bears *Cytophaga*-like 16S rDNA that places it in a separate cluster from the Sargasso Sea *Cytophaga*-like bacteria with CelM. Phylogenetic analysis of CelM genes from cultivated and uncultivated microbes revealed that the marine environmental and the *Cytophaga hutchinsonii* genes cluster together and are distinct from the CelM genes identified in a wide variety of microbes, including Archaea and Gram-positive bacteria. Sequence analysis of the Arctic CelM revealed the expected conserved domains, including COG1363 and Pfam 05343, and evidence of a signal peptide.

The physiological role of CelM and related proteins is uncertain because there is experimental evidence for genes encoding similar proteins with endoglucanase activity in *Clostridium thermocellum* and peptidase activity in *Lactococcus lactis*. In order to test the hypothesis that CelM of *Cytophaga*-like bacteria plays a role in the consumption of high molecular weight DOM by *Cytophaga*-like bacteria, we subcloned the Arctic *Cytophaga*-like CelM gene into an expression vector for endoglucanase and peptidase activity screening. We expect that the Arctic CelM will have endoglucanase activity because it is more similar to CelM in *C. thermocellum* gene than in *L. lactis*. In addition, the Arctic *Cytophaga*-like CelM is highly similarity to the CelM gene in *C. hutchinsonii*, which is an aggressive cellulose degrader. The activity of Arctic CelM is now being characterized and the complete fosmid is being sequenced by DOE-Joint Genome Institute.

87

Data Analysis and Protein Identification Strategy for the Systems-Level Protein-Protein Interaction Networks of *Shewanella oneidensis MR-1*

Gordon A. Anderson*[1] (gordon@pnl.gov), James E. Bruce[2], Xiaoting Tang[2], Gerhard Munske[2], and Nikola Tolic[1]

[1]Pacific Northwest National Laboratory, Richland, WA and [2]Washington State University, Pullman, WA

The Protein-Protein Interaction Networks research program aims to identify proteome wide protein interaction using cross-linking and high performance mass spectrometry. Cross-linking coupled with mass spectrometry is a widely used technique in protein interaction research. This technique can present a number of informatics challenges. Techniques that involve cross-linking of the proteins and then enzymatic digestion of these proteins into complex mixtures of peptides and cross linked peptides present a significant analysis challenge. The number of possible cross-linked peptides is the square of the number of tryptic peptides in the organism under study. This problem is amplified by other factors such as incomplete digestion and posttranslational modification.

This research project has developed unique cross-linker molecules called Protein Interaction Reporters (PIRs) that contain bonds that can be cleaved with high specificity in the mass spectrometer. This allows the detection of the cross-linked peptide mass and then, after low energy CID, detection of the individual peptide masses. Using this mass information and the amino acid specificity of the cross-linker molecule peptide identification is possible in many cases without further mass spectrometry. An additional feature of this technique is the spacer chain of the cross-linker that is detected in the low energy CID spectrum. This provides a mass "signature" in the CID spectrum indicating cross-linked peptide data is present. This presentation will outline the data analysis steps required and outline an identification strategy.

High performance mass spectrometry will be used to enable high throughput identification of cross linked proteins. The cross-linked proteins will be extracted from the organism and purified. The proteins will be digested and prepared for analysis by mass spectrometry. High performance FTICR mass spectrometry will be used because of its high mass measurement accuracy. FTICR allows peptide mass measurements of 1 to 5 ppm. This mass measurement accuracy coupled with the peptide constraints due to this PIR approach will enable peptide and protein identification based on mass measurement accuracy alone.

The analysis of cross linked peptides consists of a mass spectrometry experiment that involves capturing a pre-cursor ms scan containing potential cross linked peptides. The next ms spectrum is a low energy CID of the same ions detected in the first pre-cursor spectrum. The second spectrum contains the cross-linked peptides with the cross-linker fragmented. This fragmentation occurs without affecting the peptide backbone. This pair of spectra (the pre-cursor and low energy CID) allows identification of the peptide masses that were cross linked. The low energy CID spectrum will also contain a "signature" indicating a cross linked peptide pair was in the pre-cursor spectrum. This "signature" is the mass of the cross linker's core or spacer mass. This "signature" identifies spectra that potentially will identify a cross linked peptide pair. The two spectra can then be analyzed to identify the two peptide masses. The mass of the cross-linker core plus the two cross linked peptides must

* Presenting author

equal the mass of a ion detected in the pre-cursor spectrum. This analysis step will insure peptide identification will only be attempted for masses resulting from cross linked peptides.

The next step in the process is the identification of these peptides from their accurate masses. This identification is assisted by peptide constraints imposed by the cross-linker molecule. This presentation will show the mass measurement accuracy needed for identification with and without the peptide constrains this methodology provides. This analysis is performed using the genomic data from *Shewanella oneidensis*. This data will be used to calculate all of the possible cross linkable peptides. These peptide sequences will be used to calculate the peptide masses. The resultant masses will be used to predict there uniqueness at various levels of mass measurement accuracy to determine the feasibility of identifying the peptides using mass accuracy alone.

This presentation, through simulation, demonstrates the analysis algorithms that will be used to identify cross-linked peptides and illustrate the mass spectrometry performance necessary for high throughput performance.

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

88

A Protein Interaction Reporter Strategy for Systems-Level Protein Interaction Networks of *Shewanella oneidensis MR-1*

James E. Bruce[1]* (james_bruce@wsu.edu), Xiaoting Tang[1], Harry Zhu[1], Saiful Chowdhury[1], Devi Adhikari[1], Gerhard Munske[1], Gordon A. Anderson[2], and Nikola Tolic[2]

[1]Washington State University, Pullman, WA and [2]Pacific Northwest National Laboratory, Richland, WA

A key challenge inherent in the utilization of genomic data is related to deciphering the network of protein-protein interactions that enable biological function. Chemical cross-linking has gained increasing interest as a tool for protein interaction profiling, yet the examples of successful application are relatively scarce. The difficulty inherent in mass spectral interpretation and protein identification resultant from cross-linking reaction mixtures is a significant barrier that has hindered many such efforts. The analysis of complex MS and MS/MS patterns resultant from various types of cross-linker products and multiple fragmentation pathways can present levels of complexity that preclude protein and protein-protein interaction identification. However, a general technique that can identify proteins based on a physical property common to protein-protein interactions, namely, the proximity of multiple protein species within a complex mixture, is still very desirable.

Our efforts have been devoted to the development of a novel approach for chemical cross-linking that can enable improved identification of protein interactions in complex systems. A key component of this research is the development of new compounds that can provide advanced features and additional information from cross-linking reaction mixtures. We call our approach that employs mass spectrometry-cleavable cross-linkers a "Protein Interaction Reporter" (PIR) strategy, since the fragments of the cross-linker themselves are encoded with additional information that enables improved analytical capabilities for protein interaction profiling. For example, our first-generation PIR structures have been developed with low-energy CAD cleavable bonds that, when activated, release a

* Presenting author

reporter ion of specific *m/z*. These bonds can be efficiently fragmented at energy levels that preclude fragmentation of nearly all peptide amide backbone bonds. Thus, our initial MS/MS analyses of PIR-labeled products are less congested by complex multiple fragmentation pathways that are commonly observed in most cross-linked peptide MS/MS spectra. Next-generation PIR structures employ additional fragmentation schemes and features to allow even more information to be encoded in the compound. In all cases, the measured *m/z*'s of fragment ions resultant from PIR-peptide complex activation provide information that enables improved cross-link type and cross-linked peptide identification.

This presentation will highlight our initial proof-of-principle PIR experiments that were performed with model noncovalent complexes. These applications of PIR technology showed that MS/MS data could be used to differentiate various product types from cross-linking reactions, and help pinpoint ions that are resultant from protein interactions. Since the complexity of products normally poses a significant impediment to successful cross-linker application even for model noncovalent complexes, PIR advancements allow improved capabilities for analysis of protein interactions with mass spectrometry. This approach was able to produce protein-protein interaction structural data in excellent agreement with the known X-Ray crystal structure of Ribonuclease S, our model noncovalent complex. In addition, second-generation PIR structures were synthesized to incorporate affinity capture capabilities to allow enrichment of cross-linking products from complex mixtures. This feature will allow PIR cross-linked products to be enriched from digestion mixtures of proteins. Our initial investigations with this compound demonstrated that the sites of incorporation of a biotinylated PIR were nearly identical to those of the first generation structure. Thus, the affinity label seems to pose no major limitation to the reactivity of the PIR. We have also begun application of this PIR structure to complex protein mixtures, including *Shewanella oneidensis MR-1* cell lysates in an effort to better define conditions for cross-linking studies with our compounds. The results of these initial applications, additional compound features, and our envisioned implementation of the PIR strategy with tandem accurate mass analyses for the characterization of protein interactions in *Shewanella oneidensis MR-1* will be presented.

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

Technology Development and Use

Imaging, Molecular, and Cellular Analysis

89

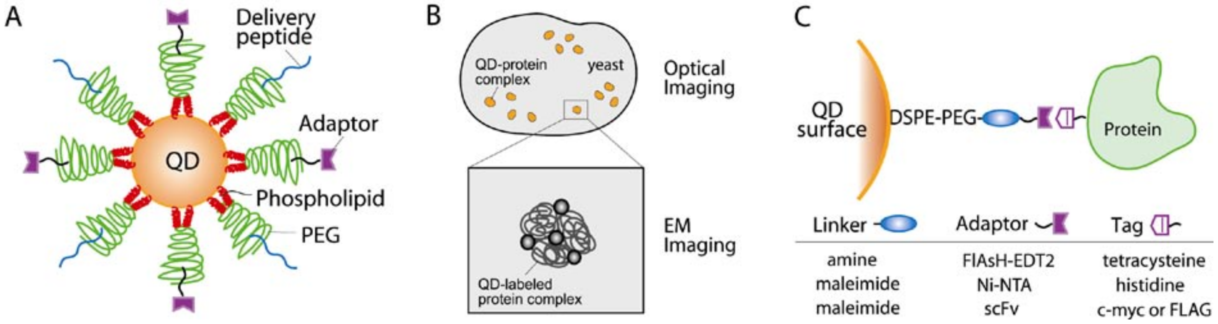
Probing Single Microbial Proteins and Multi-Protein Complexes with Bioconjugated Quantum Dots

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

[1]Georgia Institute of Technology and Emory University, Atlanta, GA; [2]California Institute of Technology, Pasadena, CA; and [3]Carnegie Mellon University, Pittsburgh, PA

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

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



* Presenting author

To achieve the goals of this DoE GTL project, we have developed quantum-dot bioconjugates with QDs encapsulated in a micelle. Phospholipids conjugated to monomethoxy PEG was used to form micelles in which the hydrophobic core of a DSPE-PEG micelle provides a cavity to encapsulate individual QDs, while the dense PEG polymer layer on the outer surface facilitates conjugation of linker molecules and delivery peptides. To facilitate bioconjugation for attaching adaptor molecules and delivery peptide, different functionalized PEG-lipid derivatives, such as DSPE-PEG-maleimide and DSPE-PEG-amine were used. To generate site-specificity, two tagging strategies were examined. The first used FlAsh-EDT$_2$ as the adaptor molecular on the QD surface and tetracysteine (Cys-Cys-Xaa-Xaa-Cys-Cys) as the tag engineered on the target protein. In the second approach, Ni-NTA was conjugated to coated QDs and the specific protein was modified to have a histidine tag. The specific targeting was demonstrated using a model system.

We performed a preliminary study of peptide-based delivery of dye molecules and quantum dot bioconjugates into yeast and *E. coli* using specifically three different peptides, TAT, polyArg, and a peptide (*ArgSerAsnAsnProPheArgAlaArg*) that has been used for delivering GFP into yeast *S. cerevisiae*. Two yeast strains, ACY 193, a wild-type yeast strain, and ACY651, a permeable yeast strain were used. We found that the polyArg peptide was the most efficient one for yeast delivery.

As part of our effort to develop QD-based technologies to identify and track individual protein complexes in microbial cells, we are advancing electron tomography as a promising new tool to image such complexes both *in vitro* and *in vivo* within small microbial cells. A new helium-cooled, 300kV, FEG, "G2 Polara" FEI TEM at Caltech was used to image purified protein complexes, viruses, and whole bacterial cells. The Polara has allowed us to record automated tilt series of a single sample cooled with either liquid nitrogen (~90K) or liquid helium (~10K). Specifically, we have recorded tilt series of purified hemocyanin and reconstructed hundreds of individual particles at various doses at each temperature, all from different "holes" of the same grid square in a single data collection session to minimize confounding variables. Surprisingly, the contrast from proteins gradually fades when they are cooled by liquid helium and iteratively imaged. Thus liquid nitrogen cooling is preferred. Using a prediction-based tracking software, we performed an automatic tilt series collection without any extra tracking or focusing images, allowing a robust data collection and reducing the time required to record a large amount of data. Further, through delivery of FEI's first "flip-flop" cryo-rotation stage, we have begun recording dual-axis tilt series of frozen-hydrated samples routinely. This has the advantages in improved point-spread-function; however, we found that our software for merging the two tilt series is not optimal, and we are presently working to improve that. These technological advances have allowed us to visualize directly cytoskeletal elements within small microbial cells and the domain structure of purified multienzyme complexes, both are key imaging goals of the Genomics:GTL program.

As a model system to study protein localization, we have been investigating the migration of *Dictyostelium discoideum* under defined extracellular stimuli. This organism responds quickly to changes in the direction of cyclic nucleotide, adenosine 3', 5'-cyclic monophosphate (cAMP) gradient. These highly polarized amoebas are characterized by continuous protrusion and retraction of pseudopodial extensions. Various localized structural responses occur in polarized *D. discoideum*. These include the localization of the β subunit of the heterotrimeric guanine nucleotide-binding proteins (G proteins) in a shallow anterior-posterior gradient, as well as the biased distribution of actin. Furthermore, the orientation of aggregating cells is dominated by the chemoattractant-induced polymerization of actin.

We have utilized custom-fabricated microfluidic devices to stimulate a cell in local domains both with two-dimensional and three-dimensional control while simultaneously visualizing its response with fluorescent microscopy using quantum dots. Both peptide-based delivery and electroporation

* Presenting author

were used to internalize QD–probes into the *Dictyostelium*. The dynamics of the actin cytoskeleton and the localization of β subunit of G protein in response to extracellular stimuli were studied using quantum-dot probes. Further, double labeling strategies were developed to localize extracellular cAMP receptors. This technique will be combined with high-resolution electron microscopy imaging to visualize individual proteins and protein complexes.

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

90

Single-Molecule Imaging of Macromolecular Dynamics in a Cell

Jamie H. D. Cate (jcate@lbl.gov) and Haw Yang* (hawyang@berkeley.edu)

Lawrence Berkeley National Laboratory, Berkeley, CA

In order to monitor macromolecular dynamics optically in living cells and to relate these observations to cellular functions two sets of tools will be essential. It will be necessary to have fluorescent probes that can be used for site-specific labeling *in vivo* and non-bleachable probes that are biocompatible with the cellular environment. The first year of this joint research project focused on the development of these two enabling technologies.

To address the need for non-bleachable and biocompatible probes, we have synthesized and characterized biocompatible gold nanoparticles, or nanotags. Nanotags will allow long-term (hours) monitoring of molecular dynamics in cells. To allow direct optical observation, the size of nanoparticles has to be greater than 10 nm. However, the surface chemistry of nanoparticles in this size range has been known to be challenging due to their propensity to aggregate. We have developed surface passivation protocols that allow stabilization of large nanoparticles and yet retain reactivity for further functionalization that is required for biological tagging.

An important aspect of these nanotags is their biocompatibility. We have developed protocols to coat the stabilized nanoparticles with various passivation agents. Biocompatibility has been stringently tested against both biochemical and biological criteria. For the former, enzymes tethered to nanotags were found to retain their reactivity, whereas for the latter, mammalian cells that contain nanotags were able to survive and propagate for several generations without aggregation of nanotags. We have also developed nanotags that will allow investigation of macromolecule rotation dynamics, as well as those that will allow multiplexing.

Finally, progress towards specific labeling of proteins with FRET donor and acceptor pairs that can be used in living cells and the status on instrumentation will be presented.

* Presenting author

91

Developing a High Resolution Method for Protein Localization in Whole Bacterium

Huilin Li* (hli@bnl.gov) and James Hainfeld (hainfeld@bnl.gov)

Brookhaven National Laboratory, Upton, NY

Bacteria lack intracellular membranes, yet the distribution and localization of many bacterial proteins are precisely controlled during cell cycle. Light microscopy has been used with great success to map the fluorescently labeled proteins, although the ~ 200 nm resolution is much to be desired, especially considering the extremely small size of the bacterial cells.

Electron tomography reaches much higher resolution than light microscopy, to ~ 10 nm, thus can be used in principle for protein localization. However one serious problem needs to be addressed. This has to do with the currently achievable resolution. Although an order of magnitude better than light microscopy, it is still short of resolving most of the proteins in the tomograms of bacteria embedded in vitreous ice. One may either strive to improve the tomogram resolution, which is the route taken by other groups, or as we decided, to specifically tag the proteins for their identification with electron dense labels, such as nanogold.

We have synthesized 3 nm diameter nanogold particles with the functionalized chemical group Ni-NTA, to be used for labeling 6X-His tagged proteins. This is being tested and optimized on expressed proteins with and without the His tag to improve specificity and efficiency of labeling.

The bacteria we choose to study, *Ralstonia metallidurans*, is 0.3-0.5 µm in thickness. To improve the contrast of the small gold label in the relatively thick transmission electron microscopy (TEM) tomograms, we will record additional scanning transmission electron microscopy (STEM) tomograms, which is known to provide enhanced contrast for high atomic number elements, such as the nanogold particles. In order to make our Jeol 2010F FasTEM/STEM microscope capable of performing tomography in both TEM and STEM modes, we have replaced the existing objective lens pole piece with a large gap one that has no limit on tilt angle. The microscope operating software FasTEM has been patched to work with the new large gap pole piece in STEM mode. We have also acquired and successfully installed a Gatan Digiscan for recording digital STEM images. The automatic tomography procedure has been implemented in TEM mode, and we are in the process of developing a Gatan Digital Micrograph-based script for automatic tomography in STEM mode. STEM tomography has been demonstrated previously in manual operation mode with plastic or inorganic samples. It is essential to record STEM tomographic tilt series in an automatic mode to minimize the radiation damage for frozen hydrated bacterial cells. We expect this to be done within a few months, and we will then proceed to image the labeled bacterial cells.

* Presenting author

92

Novel Vibrational Nanoprobes for Microbiology at the Single Cell Level

Thomas Huser* (huser1@llnl.gov), Chad E. Talley, James W. Chan, Heiko Winhold, Ted Laurence, Anthony Esposito, Christopher W. Hollars, Christine A. Hara, Allen T. Christian, Michele H. Corzett, Rod Balhorn, and Stephen M. Lane

Lawrence Livermore National Laboratory, Livermore, CA

The measurement of intracellular chemical concentrations and molecular fluxes provides essential information for systems-biological models of cells. This information, however, is difficult to obtain at the single cell level – especially in living cells where chemical levels can change rapidly in response to external or internal events.

We are studying individual microbes by a combination of optical spectroscopy techniques to obtain dynamic chemical profiles at the single cell level. Raman spectroscopy in combination with optical tweezers is used to non-destructively capture individual microbes in their native environment and assess their chemical composition within seconds. We have used this technique to dynamically monitor changes in the total protein concentration of individual cells due to increased expression after external stimulation. By focusing on changes in particular Raman peaks of a microbe we can follow trends in the overall intensity of specific peaks on an even faster timescale – down to *milliseconds* – without the need for exogenous probes. We will present examples and applications of these powerful vibrational spectroscopy techniques.

To monitor chemicals at low concentrations or chemicals that cannot typically be measured by Raman spectroscopy we also present the development of nanoscale sensors based on functionalized metal nanoparticles and surface-enhanced Raman scattering (SERS). As an example, the SERS spectrum from individual silver nanoparticle (50-80 nm in diameter) clusters functionalized with 4-mercaptobenzoic acid (4-MBA) is shown to exhibit a characteristic response to the pH of the surrounding solution, and is sensitive to pH changes in the range of 6 to 8. Measurements from nanoparticles incorporated into individual cells demonstrate that these nanoparticle sensors retain their robust signal and sensitivity to pH when incorporated into a cell. These sensors can be probed almost entirely background-free and their signals do not suffer from photobleaching, which makes them attractive long-term probes for chemical concentrations that cannot be probed by conventional Raman spectroscopy.

This work was supported by the Office of Science of the U.S. Department of Energy. Work performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract number W-7405-ENG-48.

* Presenting author

93

Instrumented Cell for Characterization of Mammalian and Microbial Cells

Jane Bearinger* (bearinger1@llnl.gov), Graham Bench, Jackie Crawford, Lawrence Dugan, Amy Hiddessen, Angela Hinz, Thomas Huser, Robin Miles, Magnus Palmblad, Chad Talley, Elizabeth Wheeler, and Allen Christian

Lawrence Livermore National Laboratory, Livermore, CA

Scientists at Lawrence Livermore National Laboratory are developing novel methodologies for high throughput cell and bacterial analysis through an internally funded Laboratory Directed Research and Development project, Instrumented Cell. The project parallels work to be conducted in the Facility for Analysis and Modeling of Cellular Systems, the final step of which requires the ability to measure and predict dynamic events within individual cells in order to achieve a comprehensive understanding of living systems. Like the Cellular Systems Facility, work focuses on experimental capabilities for the integrated observation, measurement, and analysis of the spatial and temporal variations in cellular and microbial systems.

There is a growing interest in the prospects for a new "quantitative biology", in which experimental data is integrated into predictive models of biological systems that can in turn be used to design and evaluate new experiments. There are many challenges to realizing this vision of quantitative biology; perhaps the most significant being the need to get precise data on the concentration and distribution of cellular components that will be required to develop such models. There are many analytical capabilities at LLNL that show promise for enabling quantitative measurements of cellular processes in individual cells. Instrumented Cell is developing and applying technical capabilities to measure and manipulate biochemical concentrations at the single-cell level with the ultimate goal of developing quantitative models of cellular processes. With the LLNL Microfabrication Facility, microstructures are being developed to isolate cells and make single cell measurements with the analytical tools possessed by the Chemistry and Energy and Environment directorates.

The first year of Instrumented Cell focused on production of tools capable of manipulating the environment of individual cells, and production of analytical measurements of the cells' reaction to stimuli. These capabilities provide quantitative data on an individual cell's responses to stimuli, such as characterization of uptake and effect of the new siHybrid gene silencing technology as well as the uptake of nanoparticle sensors.

Cultivation and maintenance of microbes and microbial communities under controlled conditions, including the ability to interrogate the function of individual microbial cells in the context of a characterized physicochemical environment, is another key technology needed by GTL. We are developing highly controlled systems for growing and maintaining microbial populations and communities via chemical and microfabrication-based platforms, with an emphasis on development of robust protocols that are adaptable to different strains of bacteria. We have made considerable progress in isolation of discreet bacterial communities, which will assist in high throughput analysis. Our projected plan consists of on-chip analysis of gene silencing via fluorescence experiments and off-chip cloning and mass spectrometry experiments.

* Presenting author

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48 with support from the LLNL office of Laboratory Directed Research and Development.

94

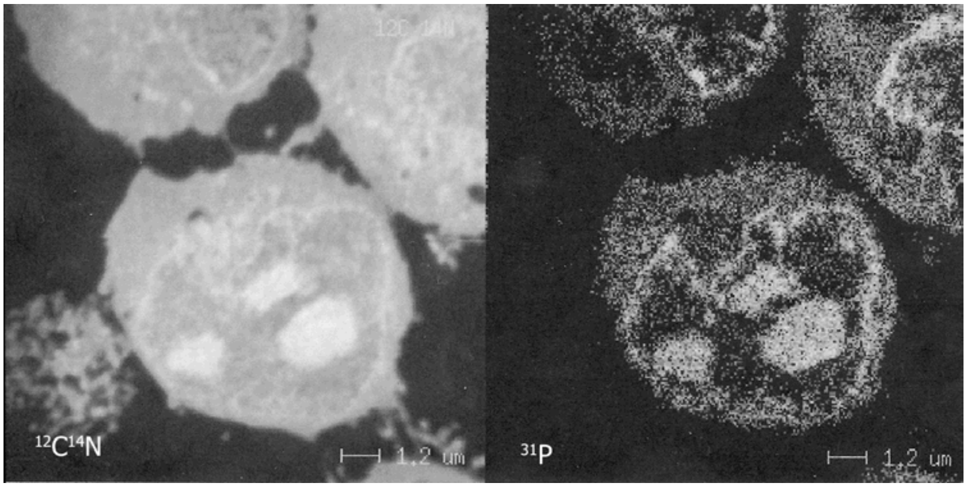
Chemical Imaging of Biological Materials by NanoSIMS

Peter K. Weber[1]* (weber21@llnl.gov), Ian D. Hutcheon[1], Radu Popa[2], and Ken Nealson[2]

[1]Lawrence Livermore National Laboratory, Livermore, CA and [2]University of Southern California, Los Angeles, CA

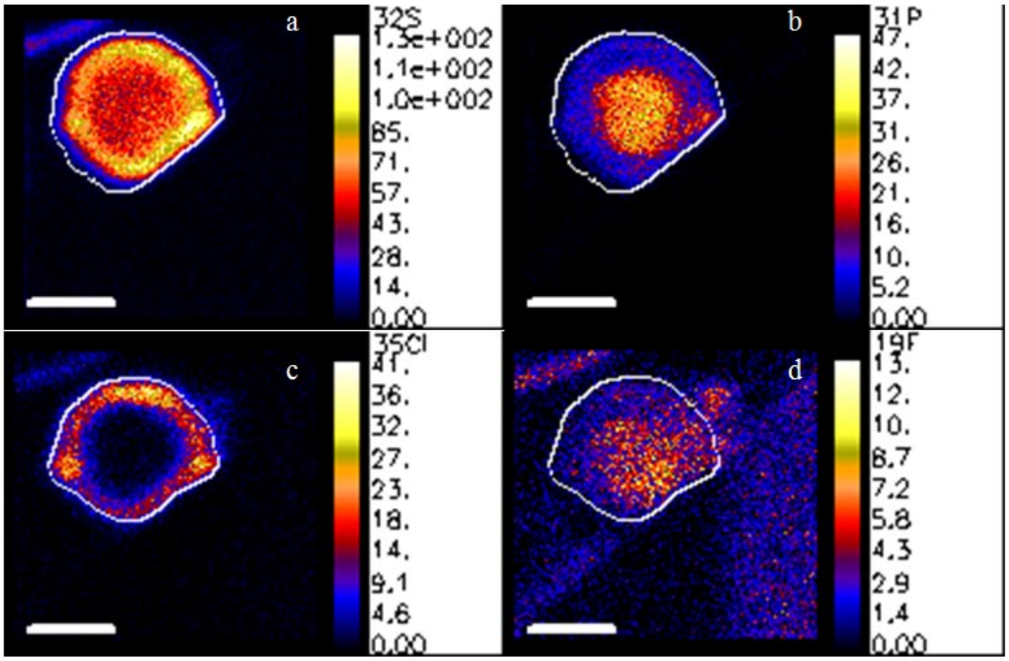
The NanoSIMS 50 represents the state-of-the-art for *in situ* microanalysis for secondary ion mass spectrometry (SIMS), combining unprecedented spatial resolution (as good as 50 nm) with ultra-high sensitivity (minimum detection limit of ~200 atoms). The NanoSIMS incorporates an array of detectors, enabling simultaneous collection of 5 species originating from the same sputtered volume of a sample. The primary ion beam (Cs^+ or O^-) can be scanned across the sample to produce quantitative secondary ion images. This capability for multiple isotope imaging with high spatial resolution is unique to the NanoSIMS and provides a novel new approach to the study of biological materials. Studies can be made of sub-regions of tissues, mammalian cells, ad bacteria. An example of the detail afforded by NanoSIMS imaging is depicted in Fig. 1, showing the distributions of N and P in individual cancer cells. Major, minor and trace element distributions can be mapped on a submicron scale, growth and metabolism can be tracked using stable isotope labels, and biogenic origin can be determined based on composition. We have applied this technique extensively to mammalian cells (Fig. 1) and bacterial spores (Fig. 2), and we are initiating a study of growth and metabolism of bacteria. Results from these studies will be discussed.

Fig. 1. NanoSIMS secondary ion images showing the distributions of N (measured as CN) and P in sectioned cancer cells.



* Presenting author

Fig. 2. NanoSIMS (a) sulfur, (b) phosphorus, (c) chlorine and (d) fluorine images of a sectioned *Bacillus thuringiensis israelensis* spore showing chemical zonation. The S and P images were collected simultaneously, and then the Cl and F images. The color bars show the total counts collected for each species, and the scale bar is 500 nm.



95

Direct Determination of Affinity in Individual Protein-Protein Complexes in Mono and Multivalent Configurations Using Dynamic Force Spectroscopy

Todd A. Sulchek[1], Kevin Langry[1], Raymond W. Friddle[1], Timothy V. Ratto[1], Sally DeNardo[2], Huguette Albrecht[2], Michael Colvin[1], and Aleksandr Noy[1,*] (noy1@llnl.gov)

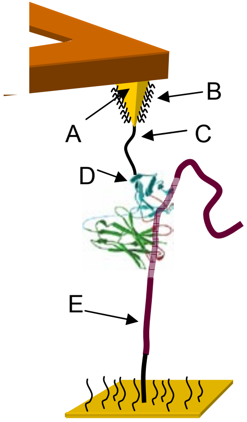
[1]Lawrence Livermore National Laboratory, Livermore, CA and [2]University of California, Davis, CA

Our laboratory at LLNL has been developing techniques for direct determination of the energy landscapes for biological molecule interactions. Interactions between proteins drive a vast variety of cellular events, and direct determination of the strength of these interactions is important to the efforts in understanding cellular metabolism and high-throughput characterization of protein complexes. Recent advances in single biological molecule manipulation and measurement have enabled direct measurements of interaction forces between individual biological molecules. We have been using atomic force microscopy (AFM) to determine energy barriers and kinetic parameters for the dissociation of individual protein-protein complexes.

* Presenting author

We used the atomic force microscope (AFM) to measure the binding forces between single molecule mucin1 (Muc1) protein and an antibody screened against Muc1. Muc1 is overexpressed on cell surfaces in a number of human cancers. Our collaborators at the UC Davis Cancer Center use antibodies to Muc1 as the targeting mechanism for delivery of radioimmunotherapeutic drugs, which consist of several such antibodies tethered to a common radioactive payload. Direct determination of binding affinities for mono and multivalent configurations of such drugs is critical for their optimization.

Figure 1: Schematic of the measurement setup. (A) gold coated tip, (B) thiol surfactant, (C,E) PEG tethers, (D) Muc1 antibody and Muc1 peptide complex.

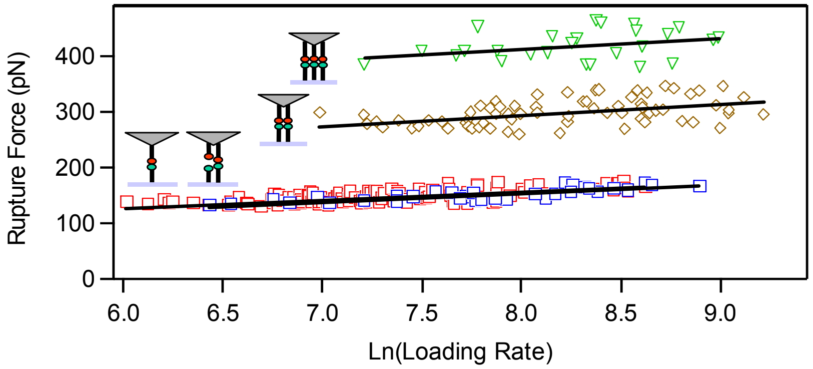


Our measurements utilized the proteins linked to the surfaces of the AFM tip and sample by flexible tethers (Figure 1). This is a versatile and general approach that spatially separates specific interactions and allows quick rejection of non-specific binding events. We have confirmed measurement of specific interactions by blocking it in a competition assay. Moreover, we were able to identify and discriminate between single and multiple rupture events by monitoring the interaction force and the nature of the tether stretch.

Measurements of the binding strength as the function of the bond loading rate (*dynamic force spectra*) allowed us to determine energy barriers, thermodynamic off-rates and the distance to the transition state for simultaneous dissociation of one, two, and three protein-protein pairs (Figure 2). Remarkably, the dynamic force spectra for single and multiple bonds show very similar slopes corresponding to the bond width for individual protein complex. These experimental observations confirm the theoretical prediction for unbinding of molecular bonds in parallel configuration. We also show that although our measured bond strength scales linearly with the number of molecule pairs, multivalent configuration leads to a precipitous decrease in the thermodynamic off-rates for the complex dissociation. Finally, we will discuss approaches for performing these measurements in high-throughput manner for potential end-line characterization of protein complexes and affinity tags.

This work was funded by the LLNL LDRD program.

Figure 2: A dynamic force spectrum showing rupture events for one (□,□), two (◊), and three (Δ) bonds. The blue square points (□) correspond to stepwise ruptures of single bonds in quick succession, and red square points (□) correspond to individual single bond rupture events. The least squares line fits predict the thermodynamic off-rates of $7\cdot10^{-3}\,s^{-1}$, $7\cdot10^{-5}\,s^{-1}$, and $4\cdot10^{-9}\,s^{-1}$ for the rupture of one, two, and three bonds respectively.



96

Electron Tomography of Intact and Sectioned Microbial Cells

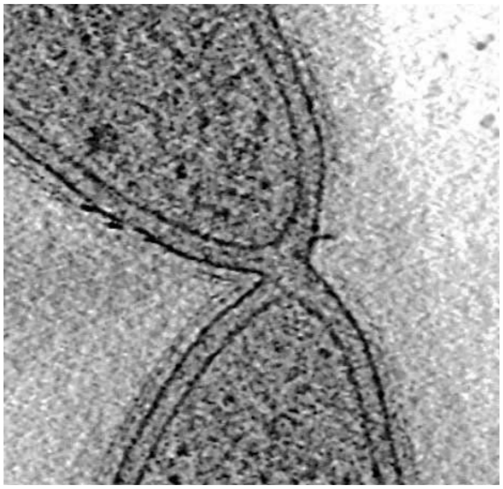
Kenneth H. Downing*[1] (khdowning@lbl.gov), Luis Comolli[1], Haixin Sui[1], Hoi-Ying Holman[1], Ellen Judd[2], and Harley McAdams[2]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA and [2]Stanford University School of Medicine, Stanford, CA

Electron tomography is an effective tool for the study of subcellular structure at a range of resolutions. In many labs tomography is being used to understand the overall structure and interplay of sub-cellular organelles of eukaryotic cells. Such work is generally carried out on plastic-embedded, stained and sectioned samples. The resolution can be high enough to identify individual molecular complexes and even to understand conformational changes associated with their functions. When cells are suitably thin, they can also be examined without sectioning. The best preservation of intact cells is obtained by rapid freezing, which forms a vitreous embedding medium that maintains the cell in a lifelike state. Although this type of preparation generally preludes the use of contrast agents, the resolution is, in principal, sufficient to identify many of the major macromolecular complexes within the cell. Such information can give insights on localization and distribution of protein complexes and will be essential for the ultimate goals of understanding and building complete computational models of the microbes.

We have been studying several microbial cells, including *Caulobacter crescentus*, *Magnetospirillum* and *Deinococcus radiodurans*, in frozen-hydrated preparations. The work with *Caulobacter* provides a particularly interesting example of the type of information one can obtain by straightforward interpretation of the 3-D data in a tomogram. The cell membranes are very well resolved in thin slices extracted from the reconstructed volume, as shown in fig. 1. Using manual or automated procedures one can segment the volume to more clearly represent features such as the membranes, as shown in fig. 2. These features could not be unambiguously interpreted from simple projection images of the cells without having seen the structure in three dimensions. This work involved examining cells close to the time at which they completed division and revealed the differential closure of the inner and outer membranes.

Figure 1. A slice 1 nm thick through the 3-D tomographic reconstruction of a *Caulobacter* cell that has almost finished dividing. The membranes are well resolved, along with parts of the periodic S-layer and subcellular densities that correspond to large protein complexes.



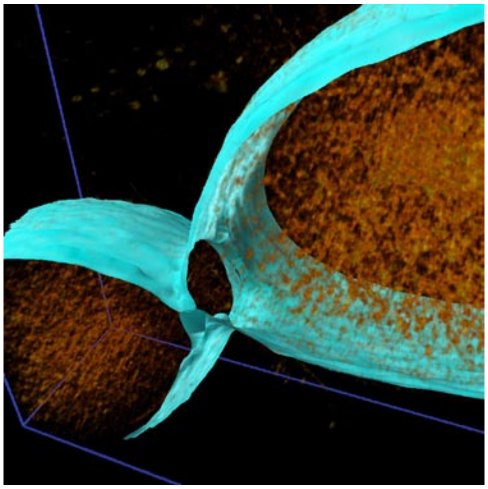
Cell membranes are among the easiest features to identify in these reconstructions because they are continuous, extended structures in three dimensions. Our goal of identifying the major macromolecular complexes will require more sophisticated template matching tools as well as the best resolution we can achieve. As a test of the achievable resolution in such work, we have been investigating the structure of a large rhabdovirus, sonchus yellow net virus. Tomographic reconstructions show

* Presenting author

very clearly the 5-nm period structure of the coiled nucleoprotein core and even the trimeric structure of the ~70 kD glycoprotein that studs the surface of the virus. This work, along with results from a number of other labs, support the projection that we will be able to identify complexes within bacteria that have a molecular weight with a lower limit of 500 – 750 kD.

In other work, we have been using conventionally embedded microbial cells to monitor changes following a change in culture conditions. *Desulfovibrio vulgaris* cells, exposed to oxygen stress, were monitored by FTIR spectroscopy. At times where the IR spectra showed interpretable changes, cells were prepared for microscopy. Changes in the 3-D structure of the cells can be correlated with apparent metabolic changes indicated by spectroscopy.

Figure 2. Surface representation of the outer membrane of a different *Caulobacter* cell in a tomographic reconstruction, also showing some of the cell contents.



97

Probing the High-Resolution Architecture and Environmental Dynamics of Microbial Surfaces by *in vitro* Atomic Force Microscopy

Alexander J. Malkin[1]* (malkin1@llnl.gov), Marco Plomp[1], Terrance J. Leighton[2], and Katherine E. Wheeler[2]

[1]Lawrence Livermore National Laboratory, Livermore, CA and [2]Children's Hospital Oakland Research Institute, Oakland, CA

The capability to image single microbial cell surfaces at nanometer scale under native conditions would profoundly impact our understanding of specific cellular processes, environmental response and bioremediation. Even though complete genome sequences are available for various microbes, the relationships between the organization and function of protein complexes within bacterial membranes and how these protein complexes respond to the change in the environment and chemical stimulants are not understood.

We have recently demonstrated that *in vitro* atomic force microscopy (AFM) can address spatially explicit bacterial spore coat protein interactions and their structural consequences at near-molecular resolution under physiological conditions. The direct visualization of the environmental response of individual *B. atrophaeus* spores revealed that upon dehydration, spore dimensions decreased by ~12%, followed by a nearly complete recovery in size upon rehydration. The observed decrease in the size of bacterial spores and concomitant change in spore coat surface morphology following dehydration are due to the contraction of the internal spore core and/or cortex. These studies establish that the dormant spore is a dynamic physical structure and provide an experimental platform for the elucidation of molecular scale bacterial spore processes, including germination, under native conditions. For the first time, species-specific high-resolution native structures of bacterial endospores including the exo-

* Presenting author

sporium and crystalline layers of the spore coat of four *Bacillus* species were visualized in their natural environment, namely air and fluid. We found that strikingly different species-dependent structures of the spore coat appear to be a consequence of nucleation and crystallization mechanisms that regulate the assembly of the outer spore coat and proposed a unifying mechanism for outer spore coat surface self-assembly.

These studies establish *in vitro* AFM as a powerful tool capable of providing a direct insight into molecular architecture and structural variability of microbial surfaces as a function of spatial, temporal, developmental and environmental organizational scales. We are currently developing approaches to utilize AFM for probing of the ultra-structure and environmental dynamics of several bacteria including *Arthrobacter oxydans* and *Thiobacillus denitrificans.*

1. M. Plomp, T.J. Leighton, K.E. Wheeler and A.J. Malkin (2005), Biophysical J., 88.

98

Real-Time Gene Expression Profiling of Single Live Cells of *Shewanella oneidensis*

X. Sunney Xie*, Jie Xiao, Ji Yu, Long Cai, Paul Choi*, Nir Friedman, Xiajia Ren, and Luying Xun*

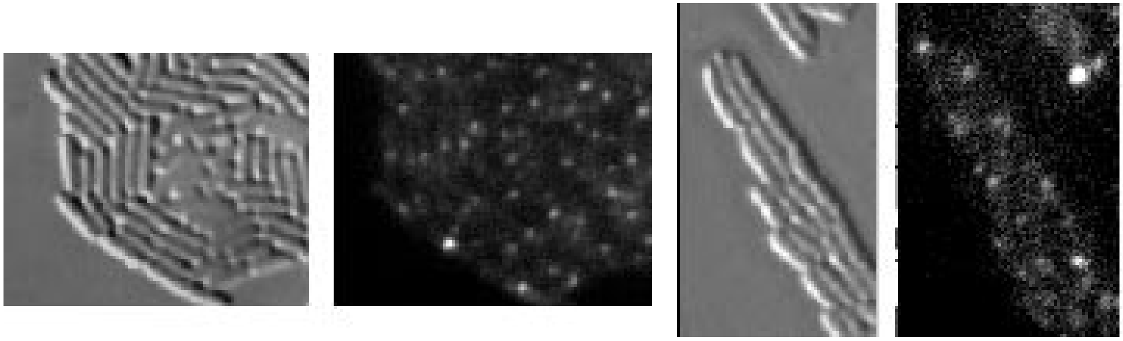
Harvard University, Cambridge, MA

Our objective is to make real-time observations of gene expression in live *Shewanella oneidensis* MR1 cells with high sensitivity and high throughput. Available technology is sufficient for the detection of gene expression at high levels; whereas, new techniques need to be developed to study expression that produces only a few protein molecules. Our efforts are divided into three areas: Developing sensitive protein reporters, adapting and fine-tuning a practical cloning strategy for library construction, and testing automation techniques for high throughput measurements with single-molecule microscopes.

Modified β-Galactosidase and green fluorescence protein (GFP) are used for reporters. Each β-galactosidase molecule generates hundreds of fluorescent molecules per second, and the product formation can be monitored in real-time. On the flipside, it is hard to observe single GFP molecules in live cells. Among the available GFP reporters, Venus GFP was chosen because of its enhanced fluorescence and short maturation time. We have detected single Venus GFP in bacterial cells after the reporter protein is attached to relatively immobile cellular components (Fig. 1). This is a major step, enabling us to follow gene expression at low signal levels. N-Terminal fusion with ubiquitin (an eukaryotic tag and degradation system) or C-terminal fusion with SsrA (a bacterial tag and degradation system) have been constructed to shorten the cellular lifetime of reporter proteins, so that cell cycle related gene expression can be monitored with single cells.

* Presenting author

Fig. 1. Differential interface contrast (left) and fluorescence (right) images of *E. coli* cells generating Venus GFP. The fusion genes are expressed at very low levels. Upon cellular immobilization, single GFP molecules are detectable as individual dots above the cell auto-fluorescence background.



A powerful cloning method that uses λ phage integrase was adapted for the construction of reporter libraries for *Shewanella oneidensis* MR-1. Cloning genes into entry vector is essentially the same as described by the commercial Gateway cloning method; however, the subsequent transfer of the cloned genes to destination vector with desired reporters is done by conjugation, an economic *in vivo* approach in *Escherichia coli*. The library is then transferred into *S. oneidensis* by conjugation. Since the destination vector cannot replicate in *S. oneidensis*, the plasmids integrate into the genome by homologous recombination between the cloned gene and the original gene on the chromosome. This creates two copies of the same gene on the chromosome separated by the reporter and vector DNA. Thus, every *S. oneidensis* gene can be tagged by a reporter to monitor gene expression. We have begun library construction of *S. oneidensis* into the entry vector. The clones can be transferred into different destination reporter vectors before being conjugated into *S. oneidensis*.

To conduct global studies with reporter libraries, we are combining microfludics with single-molecule microscopy. Microfluidic chambers, including channels and values, have been fabricated and tested with single cells carrying β-galactosidase. When fluorogenic substrate DDAO-gal is provided, the cells release the fluorescent DDAO into the chamber. Due to the fast catalysis of β-galactosidase, even a single copy of the enzyme in a cell can generate enough fluorescent signals for detection. Experiments are in progress to improve the use of microfluidics for microscopy studies and automation.

Though preliminary, our first-year results are encouraging. Complementary to DNA microarrays and mass spectrometry, our experiments will allow continuous measurements of gene expression profiling in live cells. Pushing the detection limit will enable the observation of gene expression at low signal levels, providing a complete picture of global gene expression.

99

High Throughput Fermentation and Cell Culture Device

David Klein (dklein@gener8.net), David Laidlaw, Gregory Andronaco, and Stephen Boyer

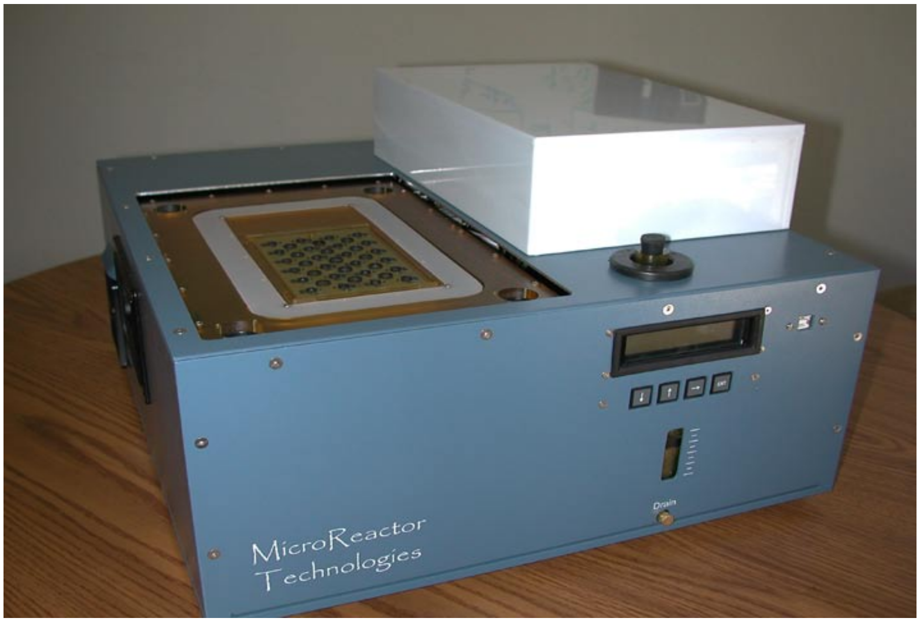
Gener8, Inc., Mountain View, CA

The Genomics:GTL (GTL) program requires that multiple microorganisms be grown with high throughput under a variety of carefully controlled-state conditions. Additionally, recombinant clones will also require culture under controlled conditions at high levels of expression, high throughput and fast production turn-around. These endeavors require technology to a) grow specific biomass under well-characterized states b) rapidly identify optimal culture conditions for expression of tagged proteins and complexes, c) rapid scale-up to obtain necessary protein samples, d) express intact protein complexes, and e) grow microbial cells in nonstandard conditions. Towards the creation of high thoughput, controlled-environment instrumentation to meet these challenges, we will present the design and function of a microreactor system with parametric controls comparable to stirred vessel bioreactors. We will further demonstrate that this type of system can be used to enhance the throughput in complex, but routine, workflows.

The system designed is a bench-top, computer-controlled microreactor system. The microreactor uses a disposable cassette (SBS standard) system with 24 individually controlled 10 ml reactors. Each reactor has independent control of temperature, pH, dissolved oxygen. The current configuration covers the following range of operation:

- Temperature (control range of 20 C to 45 C, 0.1 C resolution)
- pH (acidifying and alkalizing from pH 5 to pH 9; 0.01 resolution)
- Dissolved oxygen (0 to over 100 % air saturation, 0.1 to 10% resolution)
- Cassette agitation (0-500 RPM, 2 mm integrated orbital shaker)

Figure 1: Microreactor Prototype



* Presenting author

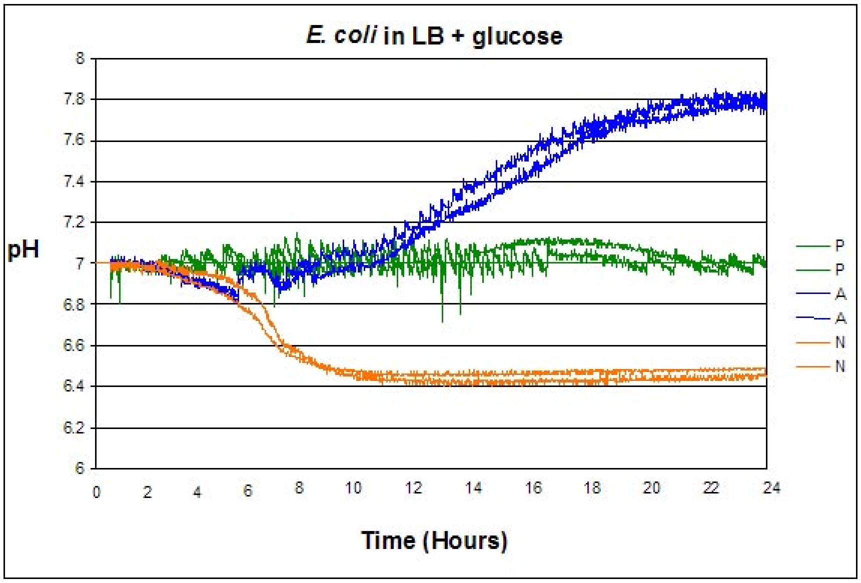
The development phase of the project has been completed. During this phase the following techniques have been developed and refined:

- Fabrication and testing of cassettes with printed sensor dots and gas permeable membranes.
- Manufacture of a 24 reactor instrumentation block with integrated heaters, gas supply manifold, and pH/DO measurement system
- Validation and calibration of the system with buffers and model system

Significant data on the use of the instrument with several biological systems of interest to GtL has been collected using the microreactor system. We will present our results on the following:

- Fine tuning of parametric control using *E. coli*, *B. subtilis and S. oneidensis*.
- Demonstration of a parametric control experiment in an organism of DOE interest; The growth and the state of the metal reduction pathway in *Shewanella oneidensis MR-1* is mapped as a function of pH, temperature, and dissolved oxygen.
- Determination of success or failure criteria on the basis of the sensor data and quality of replicates generated in the demonstration study.

Figure 2: Sample Data showing *E. coli* growth in LB with glucose. The lines marked N (orange) have neither pH control nor are supplied oxygen and thus become anaerobic and acidic. The lines marked A (blue) have oxygen control, but no pH control and thus remain aerobic and become basic. The lines marked P (green) are anaerobic, but have NH_3 based pH control enabled with a setpoint of pH 7 and so maintain a constant pH.



100

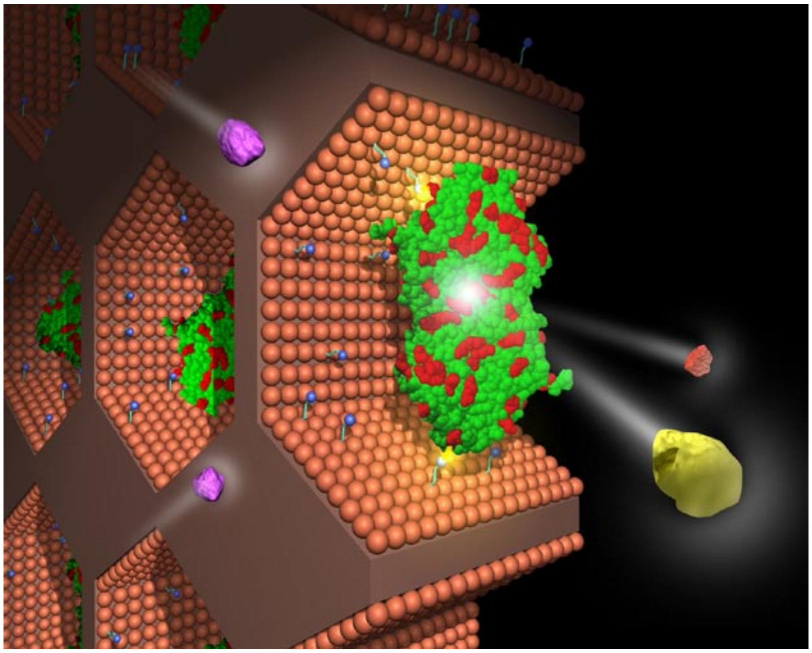
Immobilized Enzymes in Nanoporous Materials Exhibit Enhanced Stability and Activity

Chenghong Lei[1], Yongsoon Shin[1], Jun Liu[2], and Eric J. Ackerman[1]* (eric.ackerman@pnl.gov)

[1]Pacific Northwest National Laboratory, Richland, WA and [2]Sandia National Laboratories, Albuquerque, NM

Enzymes (proteins) are the nano-machines of cells. In cells, molecular crowding provides enhanced protein stability and can induce order-of -magnitude enhancements in catalytic reaction rates compared to enzymes in solution. We recently demonstrated that enzymes can be artificially crowded through immobilization on surfaces to thereby increase their reaction rates and stability. Combining appropriately functionalized, nanoporous silica (FMS) with enzymes result in immobilizations at high enzyme concentrations that exhibit enhanced stability and activity compared to enzymes in the same solution. To date we have used either carboxylethyl- or aminopropyl- FMS to either entrap or covalently immobilize three different enzymes: glucose oxidase (GOD), glucose isomerase (GI), and organophosphorus hydrolase (OPH). The working buffer and its ionic strength affected the efficiency of protein entrapment. The data is consistent with electrostatic charges contributing an important parameter governing immobilization efficiency. Net negatively charged enzymes preferred entrapment in positively-charged FMS and vice versa. The optimal percent functionalization and pore sizes must be determined empirically. In general, pore sizes slightly larger than the enzymes appear optimal. Approaches that utilize spontaneously entrapping or covalently linking with heterobifunctional crosslinking agents produce immobilized enzymes that exhibit comparable Km and Vmax to the free enzyme in solution. The combination of FMS and proteins offers and excellent platform for biological reaction engineering. Our approach could be used to make more sensitive sensors, for decontamination, to develop advanced separations based on the high specificity of protein-mediated interactions, and to generate energy enzymatically provided that suitable enzymes and proteins could be identified and produced. A potential advantage of this approach is that non-living, yet efficient enzymatic chemical reactors could be deployed in environment-friendly and environment-compatible materials (e.g. silica) without the need to maintain complex biological communities or recombinantly-engineered microbes.

Figure 1: Depicts OPH immobilized in nanoporous material reacting with substrate molecules.



* Presenting author

Protein Production and Molecular Tags

101

Towards High Throughput Selection of Binding Ligands: Using Flow Cytometry

Peter Pavlik, Milan Ovecka, Nileena Velappan, and Andrew Bradbury* (amb@lanl.gov)

Los Alamos National Laboratory, Los Alamos, NM

Phage display libraries represent a relatively easy way to generate binding ligands against a vast number of different targets. Although in principle, phage display selection should be amenable to automation, this has not yet been described and present selection protocols are far from high throughput. We have examined the selection process in a systematic approach and attempted to automate each individual step. Selection is carried out in the microtiter format using 24 targets as the individual selection lot size. Output is plated onto large assay trays, and a program to pick colonies in specific orders corresponding to the selection arrangement has been developed for the Qbot picking robot. This arrays clones according to the antigen they were selected against, and allows subsequent high density analysis using high density dot blots (up to 13,000 clones in the footprint of a microtiter plate). Although it proved possible to analyze such large numbers of clones using HD dot blots, it proved extremely difficult to quantify and digitize binding information. Furthermore, variations in expression levels led to non-specific binding artifacts – well-expressed clones gave binding signals which were often non-specific. Although this could be eliminated by also arraying clones on non-specific target filters, the integration of the information from the two filter types proved extremely difficult to quantify and analyze.

As an alternative we have examined the use of flow cytometry. In a model system, using bead based Luminex type assays, we have been able to carry out multiplex analyses, in which the reactivity of individual antibody clones for numerous different target parameters can be examined simultaneously. The analysis of each individual clone can be carried out in approximately 60 seconds, and all information is easily exportable to LIMS type systems, as well as being readily analyzed. In first experiments we were able to obtain information on the binding of individual antibody clones to specific and non-specific targets, as well as obtaining indications of expression levels. This was carried out by coupling different colored beads with: 1) specific antigen; 2) irrelevant antigen; 3) anti-tag antibodies (to determine expression level). Preliminary experiments with true selections will be presented.

* Presenting author

102

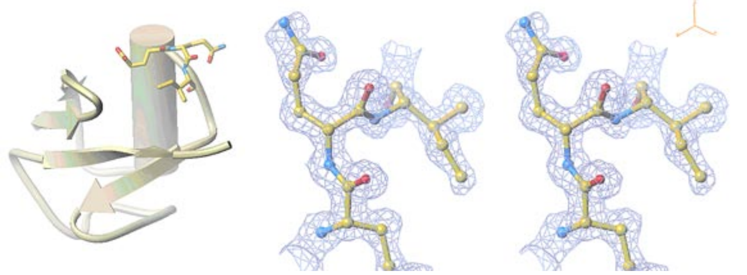
Efficient Chemical Methods for the Total Synthesis of Small Proteins: The First Crystallographic Structure of a Protein Diastereomer, [D-Gln35]-ubiquitin

Duhee Bang[1]* (duhee@uchicago.edu), George I. Makhatadze[2], and Stephen B. Kent[1] (skent@uchicago.edu)

[1]University of Chicago, Chicago, IL and [2]Pennsylvania State University, Hershey, PA

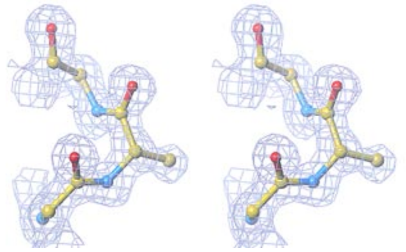
Our goal is to understand the molecular basis of the biological function of proteins using chemistry. To that end, we are developing more practical methods for the total chemical synthesis of proteins by the ligation of unprotected peptide building blocks. Our recent progress includes *a 'one pot' total synthesis of proteins* [1], *a tag assisted chemical protein synthesis* [2], and *an efficient approach to the total synthesis of cysteine free proteins* [3]. Here we will describe a case study to understand the molecular basis of protein stability using a small model protein, ubiquitin.

Figure 1. Overall fold highlighted with D-Gln35 mutation (left), 1.3 Å resolution map near the mutation (-- Glu34 – D-Gln35 – Ile36 --)



We questioned how a natural protein would adopt a D-amino acid into its overall architecture. To definitively explore changes of local and global conformations of proteins by D-amino acid incorporation, we decided to crystallize a D-amino acid incorporated protein molecule. We used a cysteine-free globular protein, ubiquitin (76 amino acids) to chemically engineer a protein α-helix. In particular, we targeted a glycine residue of the C-cap region of a protein α-helix. The conformational space of the Gly residue is only allowed for left handed α-helix and D-amino acid residues. We present (i) an efficient strategy for total chemical syntheses of ubiquitins, (ii) direct observation of the conservation of L-configuration from protein Raney-Ni reduction (Cys →Ala), (iii) the highest resolution {1.5 Å} crystal structure for known ubiquitin wild type, and (iv) high-resolution {1.3 Å} crystal structure of a ubiquitin diastereomer, UBQ[D-Gln35].

Figure 2. Conservation of L-configuration after Cys→Ala desulfurization reaction



Our syntheses made use of the native chemical ligation [4] of three unprotected peptide segments; (1-27)-thioester; (Thz28-45)-thioester; and (Cys46-76). Native Ala28 and Ala46 were replaced by Cys28 and Cys46 to enable the use of native chemical ligation at Cys. A desulfurization reaction of the product polypeptide using Raney nickel [5] was performed to convert the cysteines to alanines. Syntheses of analogue ubiquitins were performed in the same manner. The synthetic ubiquitins were crystallized and X-ray diffraction data was col-

* Presenting author

lected using the advanced photon source at ANL. We are currently refining ubiquitin wild-type and its diastereomer structures (partially refined structures are shown in Figure). The structures and ongoing efforts for the understanding of protein stability will highlight the power of the total chemical synthesis of proteins.

References

1. Bang, D. & Kent, S. B. (2004) Angew. Chem. Int. Ed. 43, 2534-2538
2. Bang, D. & Kent, S. B. *PNAS* Accepted.
3. Bang, D. G. I. Makhatadze, S. B. Kent Manuscript in preparation
4. P. E. Dawson, T. W. Muir, I. Clark-Lewis, S. B. Kent, *Science* 1994, *266*, 776-779.
5. L. Z. Yan, P. E. Dawson, *J. Am. Chem. Soc.* 2001, *123*, 526-533.

103

Development and Application of Multipurpose Affinity Probes to Isolate Intact Protein Complexes Associated with Metal Reduction from *Shewanella oneidensis* MR-1

Liang Shi*, Thomas C. Squier* (thomas.squier@pnl.gov), M. Uljana Mayer*, Haishi Cao, Baowei Chen, Yuri A. Gorby, David F. Lowry, Jeff Mclean, Seema Verma, and Ping Yan

Pacific Northwest National Laboratory, Richland, WA

Our long-term goal is to develop high-throughput methods for the rapid isolation of intact protein complexes and validation of these complexes in living cells. This methodology utilizes a small genetically encoded protein tag with an 8 amino-acid sequence containing a tetracysteine motif, which can be captured using affinity reagents or labeled with fluorescent dyes *in vivo* to permit cellular validation of protein complexes. An important advantage of this strategy is that a single small and nonperturbing tag can be sequentially used to 1) isolate the intact protein complex for identification and structural analysis and 2) visualization of the location and abundance of the protein complex within cells (Chen et al., 2004; Mayer et al., 2005). Furthermore, by varying the architecture of the affinity reagent, multiple colors and photoactivatable cross-linkers can be incorporated into the design strategy to permit measurements of binding interactions within cells and the stabilization of transient interactions associated with signaling complexes.

Proof of principle for this approach has been achieved through the isolation of two protein complexes (i.e., RNA polymerase and the metal reductase complex) from *S. oneidensis* MR-1, whose metabolism is important in understanding both microbial energy production and environmental remediation. However, these strategies will be applicable to a wide range of microorganisms and will permit the identification of environmental conditions that affect the expression of critical proteins required for the formation of transient protein complexes that facilitate bacterial growth. Our hypothesis is that identifying dynamic changes in these adaptive protein complexes will provide important insights into the metabolic regulatory strategies used by these organisms to adapt to environmental changes.

* Presenting author

RNA polymerase is a well studied system, which contains a core complex containing RNA polymerase alpha$_2$betabeta' subunits as well as regulatory proteins associated with the differential regulation of transcription. Following the expression of a tagged subunit of the RNA polymerase core complex in *S. oneidensis* MR-1, we have isolated this complex using the synthesized affinity reagent immobilized on a glass bead (Mayer et al., 2005). A critical advantage of this method is the ability to release the intact complex using a mild, one-step procedure with a competing dithiol. In addition to the identification of the core subunit complex, additional regulatory factors were identified, including the universal stress protein.

To investigate whether the current approach will also permit the identification of membrane protein complexes, we have tagged genes identified by the *Shewanella* Federation to be involved in metal reduction, and isolated members of this important protein complex. In one experiment, the metal reductase MtrC [a decaheme c-type cytochrome tentatively identified as an outer membrane protein whose activity is required for efficient reduction of Mn(IV) and Fe(III)] was genetically tagged and used to isolate two high-affinity heme-containing binding subunits in the complex that were not previously identified (i.e., OmcA and MtrA). The isolated MtrC complex maintained its activity to reduce Fe(III). This Fe(III)-reducing activity was enhanced by addition of purified MtrA, even though purified MtrA itself possessed no Fe(III)-reducing activity. Validation of this protein complex was achieved following purification of the individual proteins, using the affinity reagent dyes to measure the structural interactions between these proteins.

In summary, these multiuse affinity reagents have the advantage over other affinity tags for the high-throughput identification of protein binding partners, in that 1) the small tag can be rapidly cloned into the protein of interest and leads to minimal perturbations of binding interactions, 2) proteins are not denatured following elution permitting purification of the intact complex that can thus be further validated and studied by structural methods, and 3) the affinity reagents are cell permeable and can be used for imaging measurements to monitor protein-protein interactions in live cells.

References

1. Chen, B., M. U. Mayer, L. M. Markille, D. L. Stenoien, and T. C. Squier (2004) *Dynamic motion of helix A in the amino-terminal domain of calmodulin is stabilized upon calcium activation*. Biochemistry, in press.
2. Mayer, M. U., L. Shi, and T. C. Squier (2005) *One-step, non-denaturing isolation of an RNA Polymerase core enzyme complex using an improved multi-affinity probe resin*. J. Am. Chem. Soc., submitted.

* Presenting author

104

A Combined Informatics and Experimental Strategy for Improving Protein Expression

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

Center for Advanced Research in Biotechnology, Rockville, MD

Improved success rates for recombinant protein expression are critical to many aspects of the Genomics:GTL program.

The project is focused on determining which factors determine whether or not soluble protein is produced in *E. coli*. We are investigating the role of protein family size, native expression level, protein stability and folding rate, and the response of the host cell to expression. The outcome of the project will be a set of informatics and experimental strategies. Informatics will provide a synopsis of all relevant information for a protein, ranking alternative strategies for optimization of production. Possible new strategies include the use of reporter fusions to monitor up or down regulation of known and newly discovered cell cellular response proteins; utilization of cellular response to control cell growth; protocols for the design of mutants to improve expression; inhibition of specific proteins shown to affect outcome; and co-expression of proteins found to enhance outcome.

In the first nine months of the project, a first scan of the different factors potentially affecting expression outcome has begun. 10 proteins with representative expression properties have been prepared and submitted for micro-calorimetric investigation of their stability properties. Results are currently available for five. Messenger RNA content in *E. coli* has been investigated under conditions of overexpression of 10 proteins, five of which produce high amounts of soluble protein, and five which produce substantial amounts of insoluble material. Increased transcription of a number of genes, several of which have been implicated in stress response or protein folding, correlates strongly with the solubility status of the recombinant protein. Efforts are underway to increase or disrupt expression of those genes prior to protein induction and ascertain the effect on protein solubility.

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

* Presenting author

105

High-Throughput Production and Analyses of Purified Proteins

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

[1]Brookhaven National Laboratory, Upton, NY and [2]East Carolina University, Greenville, NC

This work is aimed at improving the efficiency of high-throughput protein production from cloned coding sequences and high-throughput biophysical characterization of the proteins obtained. Vectors and protocols for high-throughput production of proteins in the T7 expression system in *Escherichia coli* are being developed and tested by expressing and purifying proteins of *Ralstonia metallidurans*, a bacterium that tolerates high concentrations of heavy metals and has potential for bioremediation. Auto-induction allows many clones in parallel to be screened for expression and solubility simply by inoculating the cultures and growing to saturation, without the need to monitor culture growth and add inducer to each culture at the proper time. Auto-induction protocols have been developed for both BL21(DE3), in which lactose induces T7 RNA polymerase and unblocks the T7*lac* promoter, and BL21-AI, in which arabinose induces T7 RNA polymerase and lactose unblocks the T7*lac* promoter. Progress is also being made in developing new vectors that allow inducible expression of proteins that are highly toxic to the host cells. The first set of 96 *Ralstonia metallidurans* proteins is being cloned and will be tested for expression and solubility in the new vectors.

Proteins produced from clones are often improperly folded or insoluble. Many such proteins can be solubilized and properly folded, whereas others appear soluble but remain aggregated or improperly folded. As high-throughput production of purified proteins becomes implemented in GTL projects and facilities, reliable analyses of the state of purified proteins will become increasingly important for quality assurance and to contribute functional information. Beam lines at the National Synchrotron Light Source analyze proteins by small-angle X-ray scattering (SAXS) to determine size and shape, X-ray fluorescence microprobe to identify bound metals, and Fourier transform infrared (FTIR), UV circular dichroism (CD), linear dichroism (LD) and fluorescence spectroscopy to assess secondary structure and possible intermolecular orientation. A flexible liquid-handling system for automated loading of samples from 96-well plates for analysis at each of these stations has been built and is being implemented with purified proteins. When fully functional, the system will be capable of high-throughput analyses of size, shape, secondary structure and metal content of purified proteins, which will complement analyses such as gel filtration, mass spectrometry and NMR.

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

* Presenting author

106

Development of Genome-Scale Expression Methods

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

Argonne National Laboratory, Argonne, IL

Protein diversity suggests multiple expression strategies will be required to insure production of the highest possible proportion of cellular proteins. We are developing novel cellular and cell-free technologies to optimize the expression of cytoplasmic, periplasmic/secreted proteins and protein domains. These molecular tools contain elements that enable localization to appropriate cellular or extracellular compartments coupled with regulatory elements to permit control and coordination of protein expression. They also incorporate specific fusion components that promote protein stability and solubility or that facilitate detection, purification and/or protein characterization. Specific focus areas for *in* vivo expression in *E. coli* are as follows:

- Evaluation of various fusion tag cassettes to maximize the generation of soluble proteins or protein domains for downstream analysis.
- Development of a periplasmic expression system compatible with current standard high throughput cytoplasmic cloning strategies. This process has been implemented in a 96-well plate format and is being used for analysis of expression and solubility for *Shewanella* and *Geobacter* proteins directed to the cytoplasm or periplasm.
- Evaluation of a domain-based cloning and expression strategy for simple architecture membrane proteins. Proteins were analyzed for periplasmic signal sequences by sequence analysis using the signalP algorithm (1, 2) or for transmembrane regions by application of the TMHMM program (3). This approach is being applied to a set of two-component sensor and methyl accepting chemotaxis proteins from *Shewanella* and *Geobacter*.
- Generation of several constructs intended to facilitate cloning and expression of genes coding for c-type cytochromes. These constructs and host strains are being evaluated for implementation in a high throughput environment.

Our studies indicate a large fraction of proteins of highest interest are difficult to express using standard expression systems. Our novel expression methods extend the boundaries of current high throughput technology and provide strategies for expression of challenging proteins that can be implemented by the general scientific community. We are attempting to optimize distribution of purified proteins or clones that express soluble protein for characterization in detail and elucidation of biological function.

1. Nielsen, H., and Krogh, A. (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc Int Conf Intell Syst Mol Biol 6*, 122-30.
2. Nielsen, H., Brunak, S., and von Heijne, G. (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng 12*, 3-9.
3. Sonnhammer, E. L., von Heijne, G., and Krogh, A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol 6*, 175-82.

* Presenting author

107

Plate-Based Methods for Expression of Cytoplasmic Proteins from *Shewanella oneidensis*

Elizabeth Landorf[1], Terese Peppler[1], Sarah Fey[1], Alexander Iakounine[3], Eugene Kolker[2], and Frank Collart[1]* (fcollart@anl.gov)

[1]Argonne National Laboratory, Argonne, IL; [2] BIATECH, Bothell, WA ; and [3]University of Toronto, Toronto, Canada

Transcriptomic and proteomic analyses by the *Shewanella* Federation suggest there are a large number (>500) of ORFs identified as hypothetical genes that are expressed in *S. oneidensis* cells both as mRNAs and proteins. We have expressed many of these "hypothetical" proteins using plate base methods to identify strategies appropriate for genome scale analysis. Expression was confirmed for 61% of the hypothetical target set by denaturing gel electrophoresis. This represents a lower bound since not all targets were analyzed for expression. For 21% of the targets, an obvious fusion protein was not observed after gel analysis. Soluble proteins were grouped into several relative categories based on qualitative assessment of band intensity (Commassie blue staining) after denaturing gel electrophoresis. Approximately two-thirds of clones expressing fusion products generated a protein that was soluble at an analytical scale. The solubility assignments at this analytical scale show a general correlation to the yield of soluble protein obtained in preparative scale cultures suggesting most of these samples would be suitable for functional interrogation. These expression results are consistent with the genomic expression analysis indicating the assigned ORFs produce an expression product and provide an opportunity for characterization of the samples expressed in soluble form. This analysis is essential as it will provide a protein resource for the larger scientific community that enables verification of functional hypothesis derived by the various functional prediction methods. Where possible, we are attempting to experimentally validate the functional assignment. In collaboration with the University of Toronto, we have submitted a number of purified protein samples for characterization using functional screen matrix for enzymatic activity. Functional assignments were validated for several proteins in the initial test set and new samples are being prepared for submission.

* Presenting author

108

Generating scFv and Protein Scaffolds to Protein Targets

Brian K. Kay* (bkay@anl.gov), Michael Scholle, Ushma Kriplani, John Kehoe, and Frank Collart

Argonne National Laboratory, Argonne, IL

We have recently used a library of human single-chain Fragments of variable regions (scFv) to generate antibodies to a collection of *Geobacter*, *Shewanella*, and *Rhodobacter* proteins. The proteins were overexpressed in *E. coli*, chemically biotinylated, and captured on streptavidin coated magnetic beads for screening a large library of human scFv molecules, which were displayed on the surface of bacteriophage M13. After three rounds of affinity selection with liquid handling robotic workstations, the binding of isolates is confirmed by inserts are transferred *en masse* via ligation independent cloning into an alkaline phosphatase (AP) fusion vector for enzyme-linked binding assays (ELBA). The scFv-AP fusions can be used to check the specificity of the antibodies and probe western blots of cell lysates. In addition, soluble forms of the scFvs can be use to pull-down protein complexes from cell lysates and identify cellular machines. Finally, to overcome one potential limitation (i.e., presences of disulfides that may not form in reducing environments) of scFvs to perturb target functions intracellularly, we are currently exploring the use of FN3 monobodies and the villin headpiece as antibody-like scaffolds.

109

Cell Free Approaches for Protein Production

Gerald W. Becker[2]*, Pavel Shiyanov[2], Yifei Wu[2], Sarah Fey[1], Elizabeth Landorf[1], Terese Peppler[1], and Frank Collart[1] (fcollart@anl.gov)

[1]Argonne National Laboratory, Argonne, IL; and [2]Roche Applied Science, Indianapolis, IN

Cell-free protein synthesis is an easily automated screening tool and is also a scalable process for the production of preparative amounts of protein. We developed a series of plate-based methods for protein expression and solubility screening using cell-free technology that can be applied to entire genomes. In a pilot study, we selected 384 ORFs from *Shewanella oneidensis* and amplified the coding regions with KOD polymerase. The same amplification products were used for parallel expression studies using cell-free technology and *Escherichia coli* as an *in vivo* expression system. Comparison of expression/solubility outcomes indicate a general correspondence of these parameters for the two expression strategies but there are a number of targets which were only expressed and/or soluble in one of the systems. These targets are being further evaluated to determine if specific classes of protein are more readily produced using the cell-free technology. These initial studies were extended to include a series of plate-based solubility screens that evaluated the effect of various additives such as detergents, cofactors, chaperones or auxiliary proteins. Application of this approach to a set of 24 samples indicates this is an effective strategy to improve expression/solubility outcomes for many challenging proteins. The solubility enhancement screen has been implemented in a 96-well plate format and the open nature of the cell-free system will facilitate the evaluation of additional additives in the transcription/translation reaction. These preliminary studies illustrate the many advantages of

* Presenting author

cell-free expression systems such as the capability to support expression directly from PCR products without cloning, propagation and purification of plasmid DNA. This approach allows for an increase in the throughput capabilities for domain interrogation as well as enabling the screening for the suitability of different tags and fusion partners and the optimization of solubility by high throughput compatible methods.

110

Rapid Synthesis of Peptidic and Peptidomimetic Ligands for High-Throughput Protein Purification and Labeling

Jeffrey B.-H. Tok[1]* (tok2@llnl.gov), Priscilla Chan[1], David Smithson[2], Ted Tarasow[1], and Rod Balhorn[1]

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

The ability to rapidly generate protein-specific affinity reagents is clearly a desirable capability toward the goal of GTL program. My laboratory has utilized a combination of chemical and biochemical combinatorial approaches to generate such affinity tags against bacterial protein targets. Specifically, we have recently established an extremely versatile and useful approach to synthesize peptide libraries, in which several peptide libraries containing up to 2.5 million unique peptide ligands had been synthesized in ~two weeks. These libraries had subsequently been utilized to successfully identify novel affinity binders against the ganglioside-binding protein domain of the bacterium *Clostridium tetanus*. Calorimetric assays to enable rapid screening of the peptide library have also been demonstrated and have subsequently afforded "lead" sequences to enable a second generation of peptide ligands.

In addition, we have recently successfully applied the peptide affinity tags to purify the proteins through immobilizing the peptide affinity tags onto solid resins. Attempts to render this effort in a high-throughput fashion will be discussed.

* Presenting author

Proteomics and Metabolomics

111

Development and Application of New Technologies for Comprehensive and Quantitative High Throughput Microbial Proteomics

Richard D. Smith* (rds@pnl.gov), Mary S. Lipton, James K. Fredrickson, Matthew Monroe, Eric Livesay, Konstantinos Petritis, Joshua Adkins, Gordon A. Anderson, Kim Hixson, Ruihua Fang, Rui Zhao, Ronald J. Moore, and Yufeng Shen

Pacific Northwest National Laboratory, Richland, WA

With recent advances in whole genome sequencing for an increasing number of organisms, biological research is increasingly incorporating higher-level "systems" perspectives and approaches. Biology is transitioning from a largely qualitative descriptive science to a quantitative, ultimately predictive science. Key to supporting advances in microbial and other biological research at the heart of the DOE Genomics: GTL program is the ability to quantitatively measure the array of proteins (i.e., the proteome) in various biological systems under many different conditions. The challenges associated with making useful comprehensive proteomic measurements include identifying and quantifying large sets of proteins that have relative abundances spanning many orders of magnitude, which vary broadly in chemical and physical properties, have transient and low levels of modifications, and are subject to endogenous proteolytic processing. Ultimately, such measurements, and the resulting understandings of biochemical processes are expected to enable development of predictive computational models that could profoundly affect environmental clean-up and energy production by, for example, providing a more solid basis for mitigating the impacts of energy-production-related activities on the environment and human health.

In FY 2001, a project was initiated to develop quantitative and high throughput global proteomic measurement capabilities for microbial systems. The platform is based on a combination of advanced separations and mass spectrometric instrumentation and supporting computational infrastructure. The scope has included the development of an ultra-high pressure nano-scale capillary liquid chromatography platform combined with Fourier transform ion cyclotron resonance mass spectrometry and supporting data analysis and management capabilities. These developments provided the first high throughput mode "24/7" operation of such instrumentation, and resulted in its successful application to a set of microbial systems. The biological applications of this technology and associated activities are the subject of a separate, but interrelated project (J. K. Fredrickson, PI) involving a number of microbial systems (e.g. *Shewanella oneidensis* MR1, *Deinococcus radiodurans* R1*s*) in collaboration with leading experts on each organism. These studies have demonstrated the capability for automated high-confidence protein identifications, broad proteome coverage, and the capability for exploiting stable-isotope (e.g. ^{15}N) labeling methods to obtain high precision relative protein abundance measurements from microbial cultures.

A present emphasis of this project is the need for higher throughput proteomics measurements. A "prototype high throughput production" lab was established in FY 2002 was an early step in this direction. Operations within this lab are distinct from technology development efforts, both in laboratory space and staffing. This step was instituted in recognition of the different staff "mind sets"

* Presenting author

required for success in these different areas, as well as to allow "periodic upgrades" of the technology platform in a manner that does not significantly impact its production operation. The result has been faster implementation of technology advances and more robust automation of technologies that improve overall effectiveness.

Our efforts currently in progress aim to:

- Significantly increase the overall data quality of global proteome measurements, and provide data that are quantitative and have statistically sound measures of quality.
- Increase overall data production by more than an order of magnitude in conjunction with the improved data quality.
- Provide the informatics tools and infrastructure required to support improved data quality and increased throughput and to efficiently manage, use, and disseminate large quantities of data generated by GTL "users."
- Develop the foundation for the further extension of proteomics measurements to enable more comprehensive coverage of protein modifications.

A significant challenge is the immense quantities of data that must be managed and effectively managed, analyzed, and communicated with associated measures of data quality in order to be useful. Thus, a key component of our program involves the development of the informatics tools necessary to make the data more broadly available and for extracting knowledge and new biological insights from large and complex data sets.

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

112

Characterization of *Rhodobacter sphaeroides* by High Resolution Proteomic Measurements

Mary S. Lipton*[1] (Mary.Lipton@pnl.gov), Timothy Donohue*[2] (tdonohue@bact.wisc.edu), Samuel Kaplan*[3] (Samuel.Kaplan@uth.tmc.edu), Stephen Callister[1], Matthew E. Monroe[1], Margie F. Romine[1], Ruihua Fang[1], Carrie D. Goddard[1], Nikola Tolic[1], Gordon A. Anderson[1], Richard D. Smith[1], Jim K. Fredrickson[1], Miguel Dominguez[2], Christine Tavano[2], Xiaihua Zeng[3], and Jung Hyeob Roh[3]

[1]Pacific Northwest National Laboratory, Richland, WA; [2]University of Wisconsin, Madison, WI; and [3]University of Texas Medical School, Houston, TX

Exploiting microbial function for purposes of bioremediation, energy production, carbon sequestration and other missions important to the DOE requires an in-depth and systems level understanding of the molecular components of the cell that confer its function. Inherent to developing this systems level understanding is the ability to acquire global quantitative measurements of the proteome (i.e. the proteins expressed in the cell). We have applied out state of the art proteomics technologies based upon high-resolution separations combined with Fourier transform ion cyclotron resonance mass spectrometry to obtain quantitative and high throughput global proteomic measurements of

* Presenting author

the photosynthetic bacterium *Rhodobacter sphaeroides.* Significant progress has been made addressing biological questions using high resolution proteomic measurements of cells, and fractions thereof, cultivated under varying conditions.

Rhodobacter sphaeroides 2.4.1 is α-3 purple nonsulfur eubacterium with an extensive metabolic repertoire. Under anaerobic conditions, it is able to grow by photosynthesis, respiration and fermentation. Aerobically it can grow by respiration as a chemoheterotroph. It can also be grown either photo- or chemo- lithotrophically on hydrogen and carbon dioxide. The organism can fix nitrogen under anaerobic conditions, and can use a wide diversity of terminal electron acceptors such as oxygen, metal oxides or oxyanions and an array of organic molecules as electron donors. When grown photosynthetically, it uses wavelengths of light in the near infra-red and contains a reaction center that is the ancestor of plant photosystem II. *R. sphaeroides* has been shown to possess two chromosomes, the larger of approximately 3.0 Mbp and the smaller of approximately 1.0 Mbp and 5 plasmids that together encode some 4600 gene products.

The initial mass tag database consisted of global proteomic preparations from the organism cultured under both steady state aerobic and photosynthetic conditions. However, important in the physiology of the organism is not just the global expression of proteins but also the localization of these proteins. For example, the transition of the organism between an aerobic to a photosynthetic state is accompanied by a synthesis of the photosynthetic membrane imbedded with the photosynthetic apparatus. It is therefore important to determine the localization of the proteins with in the organism to achieve a clear view of the physiology. To this end, cellular fractions of these organisms cultured under both highly aerobic conditions where photosynthetic membrane synthesis is repressed (30% O_2) and photosynthetic cell states (low, $3W/m^2$, light intensity to maximize photosynthetic membrane synthesis) have been analyzed. Photosynthetic cells have been fractionated into 5 relatively discreet fractions (cytosol, periplasm, inner membrane, photosynthetic membrane and outer membrane) and the aerobic cells have been fractionated into 4 relatively discreet fractions (cytosol, periplasm, inner membrane, and outer membrane) in an effort to determine protein localization in the cell. We will report on the identification of ~XXXX total proteins from aerobic and photosynthetically-grown cells as well as the localization of proteins associated with assembly, function or control of the photosynthetic apparatus to individual subcellular fractions from steady-state photosynthetically grown cells.

The true understanding of the transition between the steady states will be achieved by a temporal study of the protein expression patterns in aerobically grown *R. sphaeroides* cells shifting to photosynthetic conditions. We have applied quantitative proteomics measurements to cells taken from a time course experiment of these cells transitioning between the two states. Preliminary studies are focused on the synthesis and deposition of the photosynthetic apparatus into the cells, however, through clustering analysis we will be able to identify other proteins that are important in this transition as well.

113

Quantitative Metalloproteomics

Patrick G. Grant* (pggrant@llnl.gov), Sharon Shields, Magnus Palmblad, and Graham Bench

Lawrence Livermore National Laboratory, Livermore, CA

Numerous bacteria have unusual enzymatic capabilities especially extreomophiles or bacteria that thrive in environments with extreme conditions (heat, acid, cold, etc…). Enzymatic activity commonly involves metal ions involved at the active site of the protein. To understand and utilize the enzymatic capability of these bacteria, enzymatic metalloproteins must be isolated, quantified, characterized, and identified as a function of the exposure to environmental cues(ions, pH, salt concentration). We have developed a nondestructive quantitative method to measure the amount of an isolated protein, the elements within that protein and identify the same protein sample with MALDI-TOF/MS by peptide mass fingerprinting.

Proteins are a critical class of biomolecule and the study of proteins, proteomics, has been enhanced with the sequencing of the genes that define proteins. However, the sequencing a species genome or even the analysis of the expression of these genes does not define the quantity of a protein within a biological partition or what is the function of the protein. The expression or production of proteins is estimated to range 8 orders of magnitude. Well-defined or isolated protein samples often contain little material. Purification of larger samples can be impractical (time or resources) or impossible (single patient), and more sensitive quantitation of small amounts of proteins is the practical solution.

It is estimated that one third of all proteins in eukaryotic species contain metal atoms with similar numbers expected in prokaryotic species, and many proteins contain molecular and elemental modifications for function and activation. Quantitation of these components within proteins has proceeded within individual protein studies instead of within groups of proteins or proteomes and is far from being completed. This neglects the relationships between the proteins and systematic changes. Accurate quantitation of proteins is also problematic because current methods utilize chemically-dependent spectroscopic analyses. The quantitative response from these analyses vary significantly even within a class of proteins. Quantitative responses are commonly linear only over limited ranges, requiring simultaneous calibrations using multiple internal or external standards. This is critical in proteomic studies of biological compartments, like a tissue, which can have up to 10000 different proteins expressed at once. We are utilizing nondestructive quantitation based on physical, rather than chemical, properties of the proteins to avoid these problems.

We utilize scanning transmission ion microscopy (STIM) to quantitate the sample, which is essentially dependent only on the amount of the analyte and is equally applicable to any isolated macromolecular sample. This technique quantitates the amount of the analyte by measuring the energy loss of a three million electron Volt (MeV) proton beam as it passes through a sample. This measurement method is independent of chemical structure and produces absolute analysis independent of standards. The sample rests on a thin, uniform substrate. The substrate allows further analysis of the same quantified sample for elemental content by particle induced X-ray emission (PIXE), identification of ligands by Time-of-flight secondary ion mass spectrometry (TOF-SIMS) or Hadamard Transform Time-of-flight mass spectrometry (HT-TOF-MS) and protein identification by Matrix Assisted Laser Desorption/ionization time-of-flight mass spectrometry (MALDI-TOF). Other surface based

* Presenting author

analytical techniques (RAMAN, UV/Vis, FTIR, etc…) and quantitation of isotopically labeled ligands with accelerator mass spectrometry (AMS) is also applicable to this same sample.

PIXE is an x-ray fluorescence technique that uses the same MeV proton beam to interrogate elemental composition within specimens. PIXE provides accurate quantitation, simultaneous multi-element detection for elements with atomic number greater than 12 and is capable of micron scale spatial resolution with 0.1 mg/kg elemental sensitivity.

We have developed methods for quantifying the mass of separated proteins at high femtomole levels, metal contents of the same protein sample to low femtomole sensitivity, identification to the femtomole level, and bound labeled ligands to the attomole level. Serial analysis of an individual sample allows the accurate determination of stoichiometric relationships of the protein to bound metals, many post-translational modifications, and bound ligands without the added error of duplicating samples for each analysis method. It is then critical that the protein that is quantitated and characterized must be identified which is possible since STIM/PIXE is nondestructive.

These measurements can be coupled to capillary or nanoscale liquid separation methods such as chromatography or capillary electrophoresis through the use of fraction collection onto our sample surface and volatile buffers. We have successfully deposited proteins from chromatography systems in spots as small as few tens of micrometers. This increases the range of mass quantitation down to a few nanograms with STIM. This project was supported by Laboratory Directed Research and Development funds.

114

New Technologies for Metabolomics

Jay D. Keasling* (jdkeasling@lbl.gov), Carolyn Bertozzi, Julie Leary, Michael Marletta, and David Wemmer

Lawrence Berkeley National Laboratory, Berkeley, CA

Microorganisms have evolved complex metabolic pathways that enable them to mobilize nutrients from their local environment and detoxify those substances that are detrimental to their survival. Metals and actinides, both of which are toxic to microorganisms and are frequent contaminants at a number of DOE sites, can be immobilized and therefore detoxified by precipitation with cellular metabolites or by reduction using cellular respiration, both of which are highly dependent on cellular metabolism. Improvements in metal/actinide precipitation or reduction require a thorough understanding of cellular metabolism to identify limitations in metabolic pathways. Since the locations of bottlenecks in metabolism may not be intuitively evident, it is important to have as complete a survey of cellular metabolism as possible. Unlike recent developments in transcript and protein profiling, there are no methods widely available to survey large numbers of cellular metabolites and their turnover rates simultaneously. The system-wide analysis of an organism's metabolite profile, also known as "metabolomics", is therefore an important goal for understanding how organisms respond to environmental stress and evolve to survive in new situations, in determining the fate of metals and actinides in the environment, and in engineering or stimulating microorganisms to immobilize these contaminants.

* Presenting author

The goals of this project are to develop methods for profiling metabolites and metabolic fluxes in microorganisms and to develop strategies for perturbing metabolite levels and fluxes in order to study the influence of changes in metabolism on cellular function. We will focus our efforts on two microorganisms of interest to DOE, *Shewanella oneidensis* and *Geobacter metallireducens*, and the effect of various electron acceptors on growth and metabolism. Specifically, we will (1) develop new methods and use established methods to identify as many intracellular metabolites as possible and measure their levels in the presence of various electron acceptors; (2) develop new methods and use established methods to quantify fluxes through key metabolic pathways in the presence of various electron acceptors and in response to changes in electron acceptors; (3) perturb central metabolism by deleting key genes involved in respiration and control of metabolism or by the addition of polyamides to specifically inhibit expression of metabolic genes and then measure the effect on metabolite levels and fluxes using the methods developed above; and (4) integrate the metabolite and metabolic flux data with information from the annotated genome in order to better predict the effects environmental changes on metal and actinide reduction.

Recently, microorganisms have been explored for metal and actinide precipitation by secretion of cellular metabolites that will form strong complexes or by reduction of the metal/actinide. A complete survey of metabolism in organisms responsible for metal and actinide remediation, parallel to efforts currently underway to characterize the transcript and protein profiles in these microorganisms, would allow one to identify rate limiting steps and overcome bottlenecks that limit the rate of precipitation/reduction.

Not only will these methods be useful for bioremediation, they will also be useful for improving the conversion of plentiful renewable resources to fossil fuel replacements, a key DOE mission. For example, the conversion of cellulosic material to ethanol is limited by inefficient use of carbohydrates by the ethanol producer. Identification of limitations in cellulose metabolism and in products other than ethanol that are produced during carbohydrate oxidation could lead to more efficient organisms or routes for ethanol production – metabolomics is the key profile to identify these rate-limiting steps.

115

Characterization of Metal Reducing Microbial Systems by High Resolution Proteomic Measurements

Mary S. Lipton[1]* (Mary.Lipton@pnl.gov), Ruihua Fang[1], Dwayne A. Elias[1], Margie F. Romine[1], Alex Beliaev[1], Matthew E. Monroe[1], Kim K. Hixson[1], Yuri A. Gorby[1], Ljiljana Pasa-Tolic[1], Heather M. Mottaz[1], Gordon A. Anderson[1], Richard D. Smith[1], Jim K. Fredrickson[1], Derek Lovley[2], and Yanhuai R. Ding[2]

[1]Pacific Northwest National Laboratory, Richland, WA and [2]University of Massachusetts, Amherst, MA

Exploiting microbial function for purposes of bioremediation, energy production, carbon sequestration and other missions important to the U.S. Department of Energy (DOE) requires an in-depth and systems level understanding of the molecular components of the cell that confer its function. Inherent to developing this systems-level understanding is the ability to acquire global quantitative

* Presenting author

measurements of the proteome (i.e., the proteins expressed in the cell). We have obtained these types of measurements in a high throughput manner for the metal reducing bacteria *Shewanella oneidensis* and *Geobacter sulfurreducens* by application of our state of the art proteomics technologies based upon high-resolution separations combined with Fourier transform ion cyclotron resonance mass spectrometry. *S. oneidensis* MR-1, a Gram-negative, facultative anaerobe and respiratory generalist, is of interest to the DOE because it can oxidize organic matter by using metals such as Fe(III) or Mn(III,IV) as electron acceptors. This bacterium can also reduce soluble U(VI) to the insoluble U(IV) form, which prevents further U mobility in groundwater and subsequent contamination of down-gradient water resources. *Geobacter sulfurreducens* also is a dissimilatory metal-reducing bacterium that can reduce soluble U(VI) to insoluble U(IV). Such microbial reduction shows significant promise for *in situ* bioremediation of subsurface environments contaminated with U, Tc, and other toxic metals such as chromium.

In collaboration with the *Shewanella* Federation, we have characterized global cellular responses to changes in electron acceptors in *S. oneidensis* MR-1 cultures from a broad range of growth conditions, specifically the presence and absence of oxygen. This in-depth characterization requires not only a qualitative survey of the protein expression patterns, but also an understanding of how the levels of expression change with culture condition. To this end, we have applied quantitative proteomics approaches to characterize protein expression profiles in *Shewanella*. Relative changes in protein abundance were determined by both $^{14}N/^{15}N$ and absolute peak intensity. Each method of protein quantitation was applied to cells grown aerobically and anaerobically to determine the proteins important for electron transport within the cell. Proteins involved in the electron transport chain, as well as Fe(III) reduction were observed to increase in expression under anaerobic conditions, while proteins involved in pyruvate and malate synthesis were observed to decrease in expression under anaerobic conditions. All hypothetical and conserved hypothetical proteins are under evaluation for expression under these conditions, as well.

Other projects under the DOE Microbial Genome Program have already sequenced the *G. sulfurreducens* genome and initiated a functional genomics study to elucidate genes of unknown function in this organism. Proteomic efforts with this microorganism that complement this work have focused on creating a database of characteristic peptide mass and elution time tags, which serve as a unique 2D markers for subsequent peptide identifications. Initial global protein expression determinations have shown protein expression in most functional categories as assigned by TIGR. Thus far, quantitative analyses of protein expression patterns in *Geobacter* have been derived only by the absolute peak intensity method. Characterization of the multiple cytochrome proteins in the organism has shown interesting changes in these proteins when they are exposed to different electron acceptors. Additionally, extension of these studies to the clustering of the protein expression patterns is revealing interesting trends in proteins expression as a result of the variation in solubility among the electron acceptor.

The accuracy and precision of making these proteomic measurements is intricately linked to the analytical instrumentation, as well as to the efficiency of the sample processing methods. Advances in automation of sample processing will reduce variation among digested samples. Additionally, improved methods for quantitation and the application of increasingly sophisticated bioinformatics tools for data analysis will greatly improve the types and quality of the proteomic data available in the future.

116

Protein Complexes and Pathways

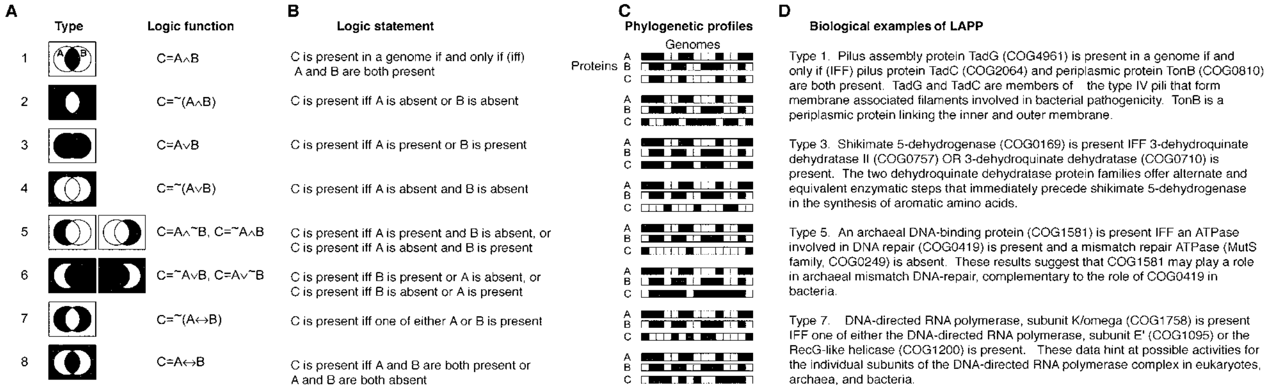
David Eisenberg* (david@mbiucla.edu), Peter Bowers, Michael Strong, Huiying Li, Lukasz Salwinski, Robert Riley, Richard Llwellyn, Einat Sprinzak, Debnath Pal, and Todd Yeates

University of California, Los Angeles, CA

Protein interactions control the life and death of cells, yet we are only beginning to appreciate the nature and complexity of their networks. We have taken several approaches towards mapping these networks. The first is the synthesis of information from fully sequenced genomes into knowledge about the network of functional interactions of proteins in cells. We analyze genomes using the Rosetta Stone, Phylogenetic Profile, Gene Neighbor, Operon methods to determine a genome-wide functional linkage map. This map is more readily interpreted when clustered, revealing groups of proteins participating in a variety of pathways and complexes. Parallel pathways and clusters are also revealed, in which different sets of enzymes operate on different substrates or with different cofactors.

These methods have been applied genome-wide to *Mycobacterium tuberculosis* and *R. Palustris*, as well as to more than 160 other genomes. Many results are available at: http://doe-mbi.ucla.edu/pronav The outcome is increased understanding of the network of interacting proteins, and enhanced knowledge of the contextual function of proteins. The information can be applied in structural genomics to find protein partners which can be co-expressed and co-crystallized to give structures of complexes.

In recent work, phylogenetic profiles have been extended by logical analysis of triplets of profiles, which reveal gene-encoded proteins that are involved in converging and parallel pathways, as well as linear metabolic pathways and complexes. The sorts of relationships uncovered are illustrated in the figure below. One example is that protein C may be present in a genome only if proteins A and B are both present. This type of logical analysis reveals many previously unidentified relationships in cellular networks because of branching and alternate pathways. It also facilitates assignment of cellular functions to uncharacterized proteins, and facilities mapping out of protein networks.



Type	Logic function	Logic statement
1	$C=A\wedge B$	C is present in a genome if and only if (iff) A and B are both present
2	$C=\sim(A\wedge B)$	C is present iff A is absent or B is absent
3	$C=A\vee B$	C is present iff A is present or B is present
4	$C=\sim(A\vee B)$	C is present iff A is absent and B is absent
5	$C=A\wedge\sim B$, $C=\sim A\wedge B$	C is present iff A is present and B is absent, or C is present iff A is absent and B is present
6	$C=\sim A\vee B$, $C=A\vee\sim B$	C is present iff B is present or A is absent, or C is present iff B is absent or A is present
7	$C=\sim(A\leftrightarrow B)$	C is present iff one of either A or B is present
8	$C=A\leftrightarrow B$	C is present iff A and B are both present or A and B are both absent

Biological examples of LAPP

Type 1. Pilus assembly protein TadG (COG4961) is present in a genome if and only if (IFF) pilus protein TadC (COG2064) and periplasmic protein TonB (COG0810) are both present. TadG and TadC are members of the type IV pili that form membrane associated filaments involved in bacterial pathogenicity. TonB is a periplasmic protein linking the inner and outer membrane.

Type 3. Shikimate 5-dehydrogenase (COG0169) is present IFF 3-dehydroquinate dehydratase II (COG0757) OR 3-dehydroquinate dehydratase (COG0710) is present. The two dehydroquinate dehydratase protein families offer alternate and equivalent enzymatic steps that immediately precede shikimate 5-dehydrogenase in the synthesis of aromatic amino acids.

Type 5. An archaeal DNA-binding protein (COG1581) is present IFF an ATPase involved in DNA repair (COG0419) is present and a mismatch repair ATPase (MutS family, COG0249) is absent. These results suggest that COG1581 may play a role in archaeal mismatch DNA-repair, complementary to the role of COG0419 in bacteria.

Type 7. DNA-directed RNA polymerase, subunit K/omega (COG1758) is present IFF one of either the DNA-directed RNA polymerase, subunit E' (COG1095) or the RecG-like helicase (COG1200) is present. These data hint at possible activities for the individual subunits of the DNA-directed RNA polymerase complex in eukaryotes, archaea, and bacteria.

These inferred interactions can be compared to directly measured protein interactions, collected in the Database of Interacting Proteins: http://dip.doe-mbi.ucla.edu/.

* Presenting author

References

1. Visualization and interpretation of protein networks in *Mycobacterium tuberculosis* based on hierarchical clustering of genome-wide functional linkage maps. M. Strong, T.G. Graeber, M. Beeby, M. Pellegrini, M.J. Thompson, T.O. Yeates, & D. Eisenberg (2003). *Nucleic Acids Research*, **31**, 7099-7109 (2003).
2. Prolinks: a database of protein functional linkages derived from coevolution. P.M. Bowers, M. Pellegrini, M.J. Thompson, J. Fierro, T.O. Yeates, and D. Eisenberg (2004). *Genome Biology*, **5**:R35.
3. Use of Logic Relationships to Decipher Protein Network Organization. P.M. Bowers, S.J. Cokus, D. Eisenberg & T.O. Yeates, *Science*, **306**, 2246-2249 (2004).

117

Metabolomic Functional Analysis of Bacterial Genomes

Clifford J. Unkefer* (cju@lanl.gov)

Los Alamos National Laboratory, Los Alamos, NM

Parallel with the terms genome, transcriptome and proteome, the combined profile of cellular metabolites is the metabolome. Examining changes in the metabolome is a potentially powerful approach to assessing gene function and contribution to phenotype. Achieving the GTL goal of obtaining a complete understanding of cellular function will require an integrated experimental and computational analysis of genome, transcriptome, proteome as well as the metabolome. Moreover, metabolites and their concentrations are a product of cellular regulatory processes, and thus the metabolome provides a clear window into the functioning of the genome and proteome. The profile of metabolites also reflects the response of biological systems to genetic or environmental changes. In addition, metabolites are the effectors that regulate gene expression and enzyme activity. *The focus of this project is the elucidation of gene function by analysis of the metabolome.* We will carry out functional studies using stable isotope labeling and Mass or NMR spectral analysis of low-molecular weight metabolites. Like the proteome, metabolic flux and metabolite concentrations change with the physiological state of the cell. Because metabolite flux and concentration are correlated with the physiological state, they can be used to probe regulatory networks. In prokaryotic organisms, the combination of functional information derived from metabolic flux analysis with gene and protein expression data being developed in other laboratories will provide a powerful approach in identifying gene function and regulatory networks. Our pilot studies will build upon our capability, demonstrate the scientific value, and establish a facility for isotope-enhanced high throughput metabolome analysis of sequenced environmental microbes.

The power of metabolome analysis will be greatly enhanced by applying the combination of stable isotope labeling and mutations. Stable Isotope labeling and NMR/Mass spectral analysis of metabolites will be used to assign metabolic function in three ways. First, we will apply specifically labeled compounds to establish precursor product relationships, and test if putative pathways identified from analysis of the genome are operational. Next, we will develop the capability for functional genomic analysis using comparative metabolomics to reveal the phenotype of a set of so-called silent mutations. This method combines null mutants constructed from the genome sequence by allelic exchange with metabolomic analysis to elucidate the function of unknown ORF's. Finally, we will carry out a full metabolic flux analysis in steady state cultures. Flux analysis will provide input for a

* Presenting author

stoichiometric model. Many of the advantages of isotope labeling for metabolomics in autotrophs and methylotrophs will be demonstrated throughout this proposal. Once demonstrated, this capability will be even more powerfully applied to heterotrophic organisms growing on complex substrates. These studies will lay the foundation to take similar labeling and metabolomic strategies into the environment to study microbial communities.

118

Dynameomics: Mass Annotation of Protein Dynamics through Molecular Dynamics Simulations of Fold-Space Representatives

David A. C. Beck* (dacb@u.washington.edu), Ryan Day, Kathryn A. Scott, R. Dustin Schaeffer, Robert E. Steward, Amanda L. Jonsson, Darwin O. V. Alonso, and Valerie Daggett

University of Washington, Seattle, WA

The Protein Data Bank (PDB) has been a tremendously useful repository of experimentally derived, static protein structures that have stimulated many important scientific discoveries. While the utility of static physical representations of proteins is not in doubt, as these molecules are fluid in vivo, there is a larger universe of knowledge to be tapped regarding the dynamics of proteins. Thus, we are constructing a complementary database comprised of molecular dynamics (MD) simulation [*1*] derived structures for representatives of all protein folds. We are calling this effort 'dynameomics.' For each fold (derived from consensus between SCOP, CATH, and DALI [*2*]) a representative protein is simulated in its native (i.e., biologically relevant) state and along its complete unfolding pathway by MD, the time-dependent integration of the classical equations of motion for molecular systems. There are approximately 1130 known non-redundant folds, of which we have simulated the first 30 that represent about 50% of known proteins. With the data resulting from our large database of MD simulations, we are data-mining for patterns and general features of transition, intermediate and denatured states to improve not only our understanding of protein dynamics but structure prediction, protein-protein and protein-ligand docking algorithms. Structure prediction remains one of the most elusive goals of protein chemistry. It is necessary to successfully predict native states of proteins, in order to translate the current deluge of genomic information into a form appropriate for functional identification of proteins from their primary sequence and rapid structure / dynamics based drug design. While these specific aims represent our immediate scientific goal for the dynameomics data, we are constructing a web site (http://www.dynameomics.org) for publication of the trajectories' coordinate data as well as in-depth analyses so that others may avail themselves of the resource and initiate areas of inquiry that we cannot even begin to anticipate.

References

1. Beck, D. A. C., and Daggett, V. (2004) Methods for Molecular Dynamics Simulations of Protein Folding / Unfolding in Solution, *Methods 34*, 112-120.
2. Day, R., Beck, D. A. C., Armen, R. S., and Daggett, V. (2003) A consensus view of fold space: Combining SCOP, CATH, and the Dali Domain Dictionary, *Protein Sci 12*, 2150-2160.

* Presenting author

Ethical, Legal, and Societal Issues

119

The DNA Files®

Bari Scott* (bariscot@aol.com)

SoundVision Productions®, Berkeley, CA

SoundVision's highly acclaimed series *The DNA Files®* has twice demonstrated that complex science could be made clear and exciting to listeners with no science background. *The DNA Files 3* will continue to show the general public the importance of cutting-edge science to their everyday lives and, at the same time, expand into a larger audience of minority and rural listeners. In addition, we plan to extend the impact of *The DNA Files 3* beyond the airwaves into schools, museums, news outlets, and even the corner coffee shop through a new network of outreach services, media projects, and learning programs along with our rich and informative web site.

The DNA Files 3 will include five nationally distributed, hour-long public radio documentaries and five 5-minute features exploring the revolutionary new developments in systems biology, neurobiology, immunology and the interaction of the environment with our DNA. Program topics will include: *Ethics Beyond the Genome: Systems Biology and Nanotechnology; Beyond DNA: The RNA World and Immunology; Individualizing the Genome: Toxicogenomics and Pharmacogenomics; Our Common Genes: Bugs, Mice and the Human Body; Depression, Addiction and Our Genes: Neurogenetics.* In addition to the documentaries themselves, SoundVision has added components to *The DNA Files®* project that will promote science journalism by less-experienced science journalists and in the ethnic media; engage the public directly in workshops employing some of the science concepts covered; and contribute to science education and journalism across a broad spectrum of platforms. The project's new outreach and education services, many of which are geared towards increasing minority and rural audiences, include:

• Media support and training

With the goal of expanding and enhancing coverage that makes science clearly relevant to a diverse population, *The DNA Files 3* will make a variety of resources available for journalists and media outlets. This and related outreach will extend the documentaries' impact by encouraging local and regional initiatives and motivating additional programs and news stories outside of SoundVision.

Materials and resources: SoundVision will provide talk show discussion topics targeted to specific ethnic communities as well as general audiences; lists of experts whom reporters can use as sources for programs and follow-up news reports, and highlights from our five documentaries. We will also develop *The DNA Files Style Book,* a handbook of best practices in science reporting on genetics and molecular biology geared to both minority-owned media and public radio as a whole. In a "push" element, SoundVision will send out short news alerts to reporters and editors to identify news stories related to *The DNA Files 3* documentaries. And, working with two ethnic media consortia, we will provide targeted background material for ethnic print.

* Presenting author

Funding and support: In addition to making the above resources generally available, SoundVision will work directly with up to 20 public radio stations. Technical support and grants (from another funding source) will be made available to stations on a competitive basis for local programming and projects related to *The DNA Files 3.* Stations will be encouraged to develop community outreach strategies and to produce programs ranging from feature reports and local documentaries to call-in programs. The impact is likely to be significant: If each of the 20 stations awarded outreach grants produced just one hour of programming, that alone would quadruple the amount of on-air content resulting from *The DNA Files 3* funding. Our outreach consultants will help these stations develop community interest, build partnerships with ethnic and minority press, and work with local science and informal learning centers.

• Educational programs

The Exploratorium, the world-renowned museum of science, art, and human perception in San Francisco, has demonstrated its support for *The DNA Files®* by agreeing to create up to five hands-on workshops that can be produced at the Exploratorium and other science museums around the country. The Exploratorium also will create a *DNA Files Workshop Kit* with materials to provide hands-on learning experiences. The kit will be designed for use by other museums, schools, churches, parks and recreation departments, as well as parents who home-school their children. It will be distributed to these outlets in tandem with the airing of *The DNA Files 3* documentaries.

• Enhanced web site

Our information-packed multimedia web site will provide tool kits to help reporters, editors, museum directors, teachers, and home-schooling parents build articles and lesson plans around *The DNA Files 3* programs. Materials on the site will include original in-depth articles related to each of the five documentaries, background information and research for editors and reporters, a library of links to related web sites, and *The DNA Files® Style Book. The DNA Files 3's* improved web site will support public radio programming and museum and school programs that can stimulate public interest in science long after the series airs.

Evaluation

An independent firm specializing in educational media will evaluate *The DNA Files 3.* Its principals will conduct annual online user surveys and interview listeners to gage their understanding and retention of the project's core themes. The evaluation also will include an evaluation of our relationship with outreach stations, plus pre- and post-production focus groups with multicultural audiences, African American audiences, and high school and junior college level biology teachers.

Project History

SoundVision, a 501(c)(3) nonprofit organization, has produced two previous *The DNA Files®* radio projects and a related multimedia web site supported by major grants from a variety of sources. *The DNA Files®* has won numerous awards, including the George Foster Peabody Award and the Alfred I. duPont-Columbia University Award. National Public Radio, which aired *The DNA Files®*, will continue to air *The DNA Files 3,* hosted by NBC Dateline reporter John Hockenberry, in its member stations.

* Presenting author

120

Science Literacy Training for Public Radio Journalists

Bari Scott* (bariscot@aol.com)

SoundVision Productions®, Berkeley, California

In the genomic era, journalists bear a greater responsibility than ever to communicate science's rapid advances and their societal implications to the public. Understanding the basics about DNA and human genetics, which most journalists still are learning, are no longer enough. Now the media must grasp concepts about the regulation of gene expression, the activity of proteins, the workings of RNA and other mechanics of the cell, both in humans and other forms of life such as microbes. Advances in these areas have profound implications, and journalists are obliged to provide clear and accurate information to the public.

SoundVision Productions®, creator and facilitator of workshops for public radio journalists, will develop a new series of three, week-long science literacy training workshops and related activities to help public radio reporters and producers meet this challenge. Through a competitive process, twelve mid-career public radio producers and reporters will be selected to attend these sessions. SoundVision will select participants who represent stations and national or regional programs that reach broad audiences. We also place a high priority on including journalists from rural and minority-controlled stations and networks. The recruitment and selection process will be designed to ensure ethnic, racial, and gender diversity.

The SoundVision workshops incorporate three goals. They are designed to increase the number and the quality of science stories produced for radio; increase the number of reporters able to report competently on complex research processes, discoveries, and resulting societal implications; and ultimately, increase civic literacy and help minimize the widening knowledge gap between the scientific community and the public. We will explore the interactions of DNA, RNA, and proteins and the overall complexity of the machinery of the cell; teach reporters what scientists are discovering about the most basic elements of life; examine the characteristics of one or more nonpathogenic microbes of environmental importance; and delve into the interactions between the human genome and the environment and toxicogenomics. A key workshop focus will be ethics in the post-human-genome-project era, including new questions about the relationship between science and business, the impact of highly patented science on society, and the risks and responsibilities of attempting to manipulate life.

Each workshop will center around 15 to 20 presentations by scientists, science journalists, scientific researchers, and radio production professionals. Sessions will orient producers to basic science, focus on the craft and responsibility of science journalism, and explore techniques for presenting complex scientific content on radio. The focus on radio is particularly important given the specific production needs that distinguish radio from other media. Each workshop will include a field trip and several informal gatherings with scientists to develop relationships and learn more about their ideas and research.

In order to reach more producers, the week-long workshops will be held in key scientific and cultural centers throughout the country. The first will be held in Boston, co-hosted by WGBH-FM and The Whitehead Institute. The second will be held in San Francisco at KQED-FM, and the third will be in Austin, Texas, in cooperation with Latino USA/KUT-FM at the University of Texas.

* Presenting author

SoundVision's training methodology has had long-lasting positive effects on public radio journalists who have attended previous workshops. Even years after attending, participants from rural to large metropolitan stations report that the workshops, which were funded by the U.S. Department of Energy, still help them with their work. Producers and reporters continue to benefit from their familiarity with the basics of DNA research, an ability to identify stories that they wouldn't previously have tackled, and better skills in getting behind press releases and scientific papers to create compelling public radio features. SoundVision's innovative workshops boost participants' confidence and their ability to communicate complex and emerging scientific research accurately. We believe that their collective work throughout the country will help lead to a better public understanding of current scientific research and its social implications.

The project also includes a website that provides transcripts and selected audio from the training sessions, "tip sheets" and online resources for participants and interested users. Follow-up teleconferences will support participants in pursuing complex and rewarding science stories for their communities. In addition, if funds permit, a pilot interactive DVD with highlights of the workshop will be offered to rural and minority-controlled stations and networks. If successful, the pilot may be developed and used for other applications.

As in our previous workshop projects, a comprehensive evaluation will be conducted by Rockman et al, a well-established, San Francisco-based evaluation firm with expertise in evaluating media projects and assessing the impact of training on journalistic practice.

Although the project targets public radio producers and reporters, with slight modifications the workshop is applicable and adaptable for news directors and editors, live interview call-in hosts and their producers (since so much air time is dedicated to that format), and even television staff..

* Presenting author

Appendix 1: Attendees

As of January 18, 2005

Carl Anderson
Brookhaven National Laboratory
cwa@bnl.gov

Gordon Anderson
Pacific Northwest National Laboratory
gordon@pnnl.gov

Adam Arkin
Lawrence Berkeley National Laboratory
aparkin@lbl.gov

Holly Baden-Tillson
J. Craig Venter Institute
HBaden-Tillson@venterinstitute.org

Ray Bair
Argonne National Laboratory
bair@mcs.anl.gov

Kristin Balder-Froid
Lawrence Berkeley National Laboratory
KHBalder-Froid@lbl.gov

Duhee Bang
University of Chicago
duhee@uchicago.edu

Gang Bao
Georgia Institute of Technology /
Emory University
gang.bao@bme.gatech.edu

Christian Barnes
J. Craig Venter Institute
cbarnes@venterinstitute.org

Paul Bayer
U.S. Department of Energy
paul.bayer@science.doe.gov

Jane Bearinger
Lawrence Livermore National Laboratory
bearinger1@llnl.gov

David Beck
University of Washington
dacb@u.washington.edu

Alexander Beliaev
Pacific Northwest National Laboratory
alex.beliaev@pnl.gov

Andrea Belgrano
National Center for Genome Resources
ab@ncgr.org

Harvey Bolton
Pacific Northwest National Laboratory
harvey.bolton@pnl.gov

Jennifer Bownas
Oak Ridge National Laboratory
bownasjl@ornl.gov

Timothy Boyle
Sandia National Laboratories
tjboyle@Sandia.gov

Andrew Bradbury
Los Alamos National Laboratory
amb@lanl.gov

James Brainard
Los Alamos National Laboratory
jbrainard@lanl.gov

Joseph Breen
National Institutes of Health
jbreen@niaid.nih.gov

David Bruce
DOE Joint Genome Institute /
Los Alamos National Laboratory
dbruce@lanl.gov

James Bruce
Washington State University
james_bruce@wsu.edu

Erick Butzlaff
University of Wisconsin
admiralexe@yahoo.com

Denise Casey
Oak Ridge National Laboratory
caseydk@ornl.gov

Parag Chitnis
National Science Foundation
pchitnis@nsf.gov

Paul Choi
Harvard University
pjchoi@fas.harvard.edu

Ray-Yuan Chuang
J. Craig Venter Institute
rchuang@venterinstitute.org

Bruce Church
Gene Network Sciences
kelly@gnsbiotech.com

George Church
Harvard University /
Massachusetts Institute of Technology
g1m1c1@arep.med.harvard.edu

Dean Cole
U.S. Department of Energy
dean.cole@science.doe.gov

Frank Collart
Argonne National Laboratory
fcollart@anl.gov

Michael Colvin
University of California, Merced
mcolvin@ucmerced.edu

Sean Conlan
Wadsworth Institute
sconlan@wadsworth.org

Maddalena Coppi
University of Massachusetts, Amherst
mcoppi@microbio.umass.edu

Robert Coyne
National Science Foundation
rcoyne@nsf.gov

Paul Crozier
Sandia National Laboratories
pscrozi@sandia.gov

Barbara Culliton
Genome News Network
bjculliton@erols.com

Patrik D'haeseleer
Lawrence Livermore National Laboratory
dhaeseleer2@llnl.gov

Sacha De Carlo
University of California, Berkeley
sachadecarlo@yahoo.com

Anjali Dhiman
Gene Network Sciences
kelly@gnsbiotech.com

Timothy Donohue
University of Wisconsin, Madison
tdonohue@bact.wisc.edu

Norman Dovichi
University of Washington
dovichi@chem.washington.edu

Kenneth Downing
Lawrence Berkeley National Laboratory
khdowning@lbl.gov

Daniel Drell
U.S. Department of Energy
daniel.drell@science.doe.gov

Leland Ellis
Department of Homeland Security
leland.ellis@dhs.gov

Larry Felser
Gene Network Sciences
kelly@gnsbiotech.com

Matthew Fields
Miami University
fieldsmw@muohio.edu

Peg Folta
Lawrence Livermore National Laboratory
folta2@llnl.gov

Marvin Frazier
J. Craig Venter Institute
mfrazier@venterinstitute.org

James K. Fredrickson
Pacific Northwest National Laboratory
tara.hoyem@pnl.gov

Haichun Gao
Oak Ridge National Laboratory
hai@ornl.gov

Timothy Gardner
Boston University
tgardner@bu.edu

George Garrity
Michigan State University
garrity@msu.edu

Sara Gaucher
Sandia National Laboratories
spgauch@sandia.gov

Al Geist
Oak Ridge National Laboratory
gst@ornl.gov

Damian Gessler
National Center for Genome Resources
ddg@ncgr.org

Daniel Gibson
J. Craig Venter Institute
dgibson@venterinstitute.org

Carol Giometti
Argonne National Laboratory
csgiometti@anl.gov

Maria Giovanni
National Institutes of Health
mgiovanni@niaid.nih.gov

Stephen Giovannoni
Oregon State University
steve.giovannoni@oregonstate.edu

John Glass
J. Craig Venter Institute
john.glass@venterinstitute.org

Peter Good
National Institutes of Health
goodp@mail.nih.gov

Yuri Gorby
Pacific Northwest National Laboratory
yuri.gorby@pnl.gov

Andrey Gorin
Oak Ridge National Laboratory
agor@ornl.gov

Patrick Grant
Lawrence Livermore National Laboratory
pggrant@llnl.gov

Chris Gunter
Nature
c.gunter@naturedc.com

David Haaland
Sandia National Laboratories
dmhaala@sandia.gov

Masood Hadi
Sandia National Laboratories
mzhadi@sandia.gov

Qiang He
Oak Ridge National Laboratory
heq1@ornl.gov

Grant Heffelfinger
Sandia National Laboratories
gsheffe@sandia.gov

Robert Hettich
Oak Ridge National Laboratory
hettichrl@ornl.gov

Peter Highnam
National Institutes of Health
Highnam@NIH.gov

Roland Hirsch
U.S. Department of Energy
roland.hirsch@science.doe.gov

Dawn Holmes
University of Massachusetts, Amherst
dholmes@microbio.umass.edu

Norman Hommes
Oregon State University
hommesn@onid.orst.edu

Brian Hooker
Pacific Northwest National Laboratory
brian.hooker@pnl.gov

John Houghton
U.S. Department of Energy
john.houghton@science.doe.gov

Greg Hurst
Oak Ridge National Laboratory
hurstgb@ornl.gov

Thomas Huser
Lawrence Livermore National Laboratory
huser1@llnl.gov

Clyde Hutchison
J. Craig Venter Institute
CHutchison@venterinsitutue.org

Prabha Iyer
J. Craig Venter Institute
piyer@venterinstitute.com

Eric Jakobsson
University of Illinois, Urbana-Champaign
jake@ncsa.uiuc.edu

Barbara Jasny
Science/AAAS
bjasny@aaas.org

Grant Jensen
California Institute of Technology
jensen@caltech.edu

Matthew Kane
National Science Foundation
mkane@nsf.gov

Samuel Kaplan
University of Texas Medical School, Houston
Samuel.Kaplan@uth.tmc.edu

Arthur Katz
U.S. Department of Energy
arthur.katz@science.doe.gov

Brian Kay
Argonne National Laboratory
bkay@anl.gov

Jay Keasling
University of California, Berkeley
keasling@berkeley.edu

Martin Keller
Diversa Corporation
mkeller@diversa.com

Stephen Kennel
Oak Ridge National Laboratory
kennelsj@ornl.gov

Vladimir Kery
Pacific Northwest National Laboratory
vladimir.kery@pnl.gov

Arnold Kim
University of California, Merced
adkim@ucmerced.edu

William Kimmerly
Pacific Northwest National Laboratory
william.kimmerly@pnl.gov

David Klein
Gener8, Inc.
dklein@gener8.net

Michael Knotek
Consultant to DOE
m.knotek@verizon.net

Eugene Kolker
BIATECH
ekolker@biatech.org

David Koppenaal
Pacific Northwest National Laboratory
david.koppenaal@pnl.gov

Henrietta Kulaga
IPTO
hkulaga@snap.org

Stephen Lane
Lawrence Livermore National Laboratory
lane12@llnl.gov

Jennie Larkin
National Institutes of Health
larkinj2@nhlbi.nih.gov

Carole Lartigue
J. Craig Venter Institute
clartigue@venterinstitute.org

Charles Lawrence
Brown University
Charles_Lawrence@Brown.edu

Philip LeDuc
Carnegie Mellon University
prleduc@cmu.edu

Adam Lee
University of Maryland
adamlee@umd.edu

Kyriacos Leptos
Harvard Medical School
leptos@fas.harvard.edu

Huilin Li
Brookhaven National Laboratory
hli@bnl.gov

Xiaoxia Lin
Harvard Medical School
xiaoxia@genetics.med.harvard.edu

Derek Lovley
University of Massachusetts
dlovley@microbio.umass.edu

William MacConnell
MacConnell Research Corporation
bmacconnell@macconnell.com

Lee Makowski
Argonne National Laboratory
lmakowski@anl.gov

Alexander Malkin
Lawrence Livermore National Laboratory
malkin1@llnl.gov

Reinhold Mann
Oak Ridge National Laboratory
mannrc@ornl.gov

Betty Mansfield
Oak Ridge National Laboratory
mansfieldbk@ornl.gov

Roummel Marcia
University of Wisconsin, Madison
rmarcia@biochem.wisc.edu

Anthony Martino
Sandia National Laboratories
martino@sandia.gov

Adam Martiny
Massachusetts Institute of Technology
martiny@mit.edu

Mahir Maruf
J. Craig Venter Institute
mahir.maruf@venterinstitute.org

M. Uljana Mayer-Cumblidge
Pacific Northwest National Laboratory
uljana.mayer-cumblidge@pnl.gov

Harley McAdams
Stanford University
hmcadams@stanford.edu

Lee Ann McCue
Wadsworth Center
mccue@wadsworth.org

Gail McLean
U.S. Department of Agriculture
gmclean@csrees.usda.gov

Barbara Methe
The Institute for Genomic Research
bmethe@tigr.org

Lisa Miller
Brookhaven National Laboratory
lmiller@bnl.gov

Marissa Mills
Oak Ridge National Laboratory
millsmd@ornl.gov

Julie Mitchell
University of Wisconsin, Madison
mitchell@math.wisc.edu

Jennifer Morrell-Falvey
Oak Ridge National Laboratory
morrelljl1@ornl.gov

Sue Morss
Argonne National Laboratory
smorss@anl.gov

Vasantha Nagarajan
DuPont
vasantha.nagarajan@usa.dupont.com

Dave Nelson
Lawrence Livermore National Laboratory
nelson6@llnl.gov

Lee Newberg
Wadsworth Center
lee.newberg@wadsworth.org

Charles Ngowe
Michigan State University
ngowe@chemistry.msu.edu

Kim Nylander
Oak Ridge National Laboratory
nylanderk@ornl.gov

Yasuhiro Oda
University of Iowa
yasuhiro-oda@uiowa.edu

Susan Old
National Institutes of Health
olds@nhlbi.nih.gov

Paola Oliveri
Caltech
poliveri@caltech.edu

Paula Olsiewski
Alfred P. Sloan Foundation
olsiewski@sloan.org

Regina O'Neil
University of Massachusetts, Amherst
rtarallo@microbio.umass.edu

Mary Oster-Granite
National Institutes of Health
mo96o@nih.gov

Brian Palenik
University of California, San Diego
bpalenik@ucsd.edu

Aristides Patrinos
U.S. Department of Energy
ari.patrinos@science.doe.gov

Dale Pelletier
Oak Ridge National Laboratory
pelletierda@ornl.gov

Cynthia Pfannkoch
J. Craig Venter Institute
cpfannkoch@venterscience.org

Anne Plant
NSTC/OSTP
aplant@ostp.eop.gov

Steve Plimpton
Sandia National Laboratories
sjplimp@sandia.gov

Gemma Reguera
University of Massachusetts, Amherst
gemma_reguera@microbio.umass.edu

Mark Rintoul
Sandia National Laboratories
rintoul@sandia.gov

Margie Romine
Pacific Northwest National Laboratory
margie.romine@pnl.gov

Denise Russo
National Institutes of Health
drusso@mail.nih.gov

Herbert Sauro
Keck Graduate Insitute
hsauro@kgi.edu

Luis Sayavedra-Soto
Oregon State University
sayavedl@science.oregonstate.edu

David Schwartz
University of Wisconsin, Madison
dcschwartz@facstaff.wisc.edu

Salvatore Sechi
National Institutes of Health
Salvatore_Sechi@nih.gov

Michael Seibert
National Renewable Energy Laboratory
mike_seibert@nrel.gov

Margrethe Serres
Marine Biological Laboratory
mserres@mbl.edu

Liang Shi
Pacific Northwestern National Laboratory
Liang.Shi@pnl.gov

Michael Sinclair
Sandia National Laboratories
mbsincl@sandia.gov

Dinah Singer
National Cancer Institute
ds13j@nih.gov

Richard Smith
Pacific Northwest National Laboratory
rds@pnl.gov

Hamilton Smith
J. Craig Venter Institute
hsmith@venterinstitute.org

Harold Smith
University of Maryland
smithh@umbi.umd.edu

Thomas Squier
Pacific Northwest National Laboratory
thomas.squier@pnl.gov

Ranjan Srivastava
University of Connecticut
srivasta@engr.uconn.edu

David A. Stahl
University of Washington
dastahl@u.washington.edu

Marvin Stodolsky
U.S. Department of Energy
Marvin.Stodolsky@science.doe.gov

F. William Studier
Brookhaven National Laboratory
studier@bnl.gov

John Sutherland
Brookhaven National Laboratory
jcs@bnl.gov

F. Robert Tabita
Ohio State University
tabita.1@osu.edu

Chad Talley
Lawrence Livermore National Laboratory
talley1@llnl.gov

Theodore Tarasow
Lawrence Livermore National Laboratory
tarasow2@llnl.gov

Michael Teresinski
U.S. Department of Energy
michael.teresinski@science.doe.gov

Michael Thelen
Lawrence Livermore National Laboratory
mthelen@llnl.gov

David Thomassen
U.S. Department of Energy
david.thomassen@science.doe.gov

Janelle Thompson
Massachusetts Institute of Technology
janelle@mit.edu

Joyce Thorpe
J. Craig Venter Institute
jthorpe@venterinstitute.org

James Tiedje
Michigan State University
tiedjej@msu.edu

Jerilyn Timlin
Sandia National Laboratories
jatimli@sandia.gov

Jeffrey Tok
Lawrence Livermore National Laboratory
tok2@llnl.gov

Emily Turner
University of Washington
emilyt@u.washington.edu

Ravishankar Vallabhajosyula
Keck Graduate Institute
rrao@kgi.edu

Sanjay Vashee
J. Craig Venter Institute
svashee@venterinstitute.org

Wim Vermaas
Arizona State University
wim@asu.edu

Judy Wall
University of Missouri, Columbia
wallj@missouri.edu

Sharlene Weatherwax
U.S. Department of Energy
sharlene.weatherwax@science.doe.gov

Peter Weber
Lawrence Livermore National Laboratory
weber21@llnl.gov, mcgurn1@llnl.gov

Xueming Wei
Oregon State University
weixue@science.oregonstate.edu

Raymond Wildung
U.S. Department of Energy
raymond.wildung@science.doe.gov

H Steven Wiley
Pacific Northwest National Laboratory
steven.wiley@pnl.gov

David Wilson
Cornell University
dbw3@cornell.edu

X. Sunney Xie
Harvard University
xie@chemistry.harvard.edu

Ying Xu
University of Georgia
xyn@bmb.uga.edu

Qing Xu
J. Craig Venter Institute
qxu@tcag.org

Luying Xun
Harvard University
xun@mail.wsu.edu

Haw Yang
University of California, Berkeley
hawyang@berkeley.edu

Jane Ye
National Institutes of Health
yej@nhlbi.nih.gov

Shibu Yooseph
J. Craig Venter Institute
syooseph@venterinstitute.org

Karsten Zengler
Diversa Corporation
kzengler@diversa.com

Jizhong Zhou
Oak Ridge National Laboratory
zhouj@ornl.gov

Julie Zimmerman
Louisiana State University
jzimme5@lsu.edu

Jeremy Zucker
Harvard Medical School
zucker@research.dfci.harvard.edu

Appendix 2: Web Sites

Program Web Sites

- Genomics:GTL Web site: http://doegenomestolife.org/
- This book: http://doegenomestolife.org/pubs/2005abstracts/
- DOE Microbial Genome Program: http://microbialgenome.org/

Web Sites Listed in Abstracts

- BioWarehouse System: http://bioinformatics.ai.sri.com/biowarehouse/
- Carbon Sequestration in *Synechococcus sp.*: From Molecular Machines to Hierarchical: Modeling http://www.genomes2life.org/
- Center for Molecular and Cellular Systems: http://www.ornl.gov/GenomestoLife/
- GelBank: http://gelbank.anl.gov/
- Joint Genome Institute: http://www.jgi.doe.gov/
- LAMMPS Molecular Dynamics Simulator: http://www.cs.sandia.gov/~sjplimp/lammps.html
- Microbial Ecology, Proteogenomics & Computational Optima. DOE Genomes to Life Center at Harvard/MIT/BWH/MGH: http://arep.med.harvard.edu/DOEGTL/
- ProteomeWeb: http://proteomeweb.anl.gov/
- Synechococcus sp. WH8102 Knowledgebase: http://csbl.bmb.uga.edu/WH8102/
- Synechocystis sp. PCC 6803: http://lsweb.la.asu.edu/synechocystis/
- The SEED: An Annotation/Analysis Tool: http://theseed.uchicago.edu/FIG/index.cgi
- Virtual Institute of Microbial Stress and Survival (VIMSS): http://vimss.org

Author Index

Indexed by page number.

Indexed by page number.

Indexed by page number.

Indexed by page number.

Indexed by page number.

Indexed by page number.

Indexed by page number.

Index

Indexed by page number.

Institution Index

Indexed by page number.

Indexed by page number.

Addendum

Abstracts received after January 24, 2005

Metagenome Analysis of Contaminated Sediments at the DOE Hanford Site

Natalia Maltsev[1], Tanuja Bompada[1], Banu Gopalan[2] (agor@ornl.gov), Shu-mei Li[3], Weiwen Zhang[3], J. Chris Detter[4], Paul Richardson[4], Margie Romine[3], and **Fred Brockman[3]**

[1]Bioinformatics Group, Argonne National Laboratory; [2]Computational Biology and Bioinformatics, Pacific Northwest National Laboratory, Richland, WA; [3]Microbiology Group, Pacific Northwest National Laboratory, Richland, WA; [4]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA

Technologies are needed to make improved inferences of microbial community function from metagenome sequence. Most microbes have evolved multiple mechanisms (programs) to capture energy. The cells' immediate biochemical and geochemical environment determines the regulatory networks that are engaged to run the best program to capture energy. Bulk extraction of RNA and analysis on microarrays largely destroys the spatial and functional linkages that are the key to understanding how communities interact. Therefore, a critical need for understanding the details of how multi-species microbial communities interact at the cellular level to generate a functional output, is the ability to interrogate the microscopic spatial organization of gene expression in these multi-species microbial communities.

The most likely methods to accomplish this objective are (1) robotic single (prokaryotic) cell "picking" by laser catapulting microscopy followed by RNA extraction, whole transcriptome amplification (if required), and probing and/or sequencing; or (2) mRNA-targeted non-PCR based fluorescence in situ hybridization (mRNA-FISH).

A second critical technology is the use of high-end computing to utilize massive amounts of metagenome sequence to design optimal phylogenetically-constrained function-oriented oligonucleotide probes for both of these approaches.

The goals of this newly funded project are to:

- Develop mRNA-targeted non-PCR based fluorescence in situ hybridization (mRNA-FISH) using soluble and nanoparticle near-infrared (low noise) probes coupled to advanced microscopy able to detect a very small number of photons
- Use grid-based computational tools to analyze community metagenome data to develop hypotheses regarding the functional processes and linkages occurring in the multi-species community, and for design of a suite of phylogenetically-constrained metabolic function "signature" probes.

The results presented in this poster relate to the second goal. As an initial exercise, we are analyzing metagenome sequence produced in a previous Microbial Genome Project. That project pooled enrichments from contaminated sediments beneath a leaking high-level radioactive waste tank at the DOE Hanford Site. Because biomass levels were very low (~10^4 cells per gram), a variety of enrichments were pooled in order to have adequate DNA to construct a clone library for sequencing. Most of the enrichments produced so little biomass that DNA concentrations were inadequate (in 2002) for constructing clone libraries. In 2003, a clone library was made from an enrichment pool that had the highest amount of DNA (only 750 nanograms). Community DNA's were also extracted in 2002 from pools of enrichments derived from more contaminated sediments, and in light of recent technological advancements clone libraries are now being constructed from those DNA samples (5 - 50 nanogram amounts) by Lucigen Corporation (Madison, WI).

Here we report on the preliminary metagenome analysis of the clone library constructed in 2003. Although the initial analysis reported here is a very small amount of sequence, we plan a minimum of 100-fold higher amounts of metagenome sequence from each library currently under construction. A total of 2,887 bacterial clones were sequenced and yielded a total of 7,071 hits representing 489 EC classes and 113 KEGG maps. At least one gene was present for synthesis of 18 of the 20 amino acids. Pathways in which ten or more genes in the pathway were present include metabolism of purines, pyrimidines, aminoacyl-tRNA, glycolysis/gluconeogenesis, pyruvate, starch, glycerolipid, porphyrin, glycine/serine/threonine, arginine/proline; and phenylalanine/tyrosine/tryptophan.

Protein hits were largely consistent with amplified and sequenced 16S rDNA phylogenies from both pooled enrichments and sediments. The 16S data from pooled enrichments showed 10 genera from the Micrococcineae, Propionibacterineae, and Steptomycineae suborders within the Actinobacteria (high GC Gram positive) phylum; and one genera (*Pseudomonas*) within the gamma class of the Proteobacteria phylum). The protein hits were 68% Proteobacteria, 30% Actinobacteria, and 3% to the Bacilli and Clostridia classes of the Firmicutes (low GC Gram positive) phylum. The *Pseudomonas* species detected in the DNA and protein hits are nitrate-reducers and are rare or absent in pristine, deep subsurface sediments at the Hanford Site; however, their presence is consistent with nitrate being the predominant inorganic contaminant in the sediments.

A web site has been constructed displaying taxonomic analysis of the metagenome; views of each contig including CDS information, potential functions, and relevant metabolic pathways; metagenome metabolic reconstruction; metabolic pathways indexed by similarity to organisms; and list of KEGG map identifications linked to organisms. Visualizations of the data using PNNL-developed Biological Data Fusion and OmniViz software will be shown. Interesting findings include (1) quite low and amino acid identities (and e-scores) for hits to the Actinobacteria and Firmicutes, suggesting a relatively novel component of this metagenome in comparison to current microbial and metagenome sequence from these phyla, and (2) a strong over-representation of transmembrane transport protein hits in the metagenome sequence. The relevance of these findings is being analyzed in detail. Future work includes metagenome sequence analysis of the very highly contaminated Hanford sediments and identification of phylogenetically-constrained metabolic function “signature” probes.

SAXS/WAXS Studies of σ54-Dependent AAA+ ATPases: Insights about Signal Transduction and Motor Function.

B. Tracy Nixon[1] (btn1@psu.edu), Baoyu Chen,[1] Michaeleen Doucleff,[2] David E. Wemmer,[2,3] Timothy R. Hoover,[4] and Elena Kondrashkina.[5]

1) Biochemistry and Molecular Biology, Penn State University, University Park, PA 16802 USA; 2) Chemistry, University of California, Berkeley, CA USA; 3) Lawrence Berkeley National Lab, Berkeley, CA USA; 4) Microbiology, University of Georgia, Athens, GA USA; 5) BioCAT at APS/Argonne National Lab, Illinois Institute of Technology, 9700 South Cass Ave, Argonne, IL 60439, USA.

AAA+ ATPases are molecular motors that provide important biological functions in all kingdoms of life. We are still learning how their actions are controlled and how they perform mechanical work. The most prevalent information processing system in bacteria, two-component signal transduction, is sometimes used to regulate the assembly AAA+ ATPase machines that regulate transcription by the σ54-form of RNA polymerase. Powerful and complementary microscope technologies are being developed under the GTL project 'Microscopes of Molecular Machines (M3): Structural Dynamics of Gene Regulations in Bacteria' (C*arlos Bustamante*, PI) to study the structure and changes that occur when multi-protein molecular machines such as AAA+ ATPases are formed. The technologies (cryo-electron microscopy, atomic force microscopy, optical tweezers, and single-molecule fluorescence microscopy) look at molecular machines from different but complementary perspectives. Cryo-electron microscopy can, for example, form visual images of an entire molecular complex, while single-molecule fluorescence can show real time formation of complexes as a result of fluorescence signals that can be seen as specific proteins come in contact with each other. Although static light scattering from molecules in solution occurs from molecules in all possible orientations, shape information is available to about 5 Angstroms resolution. The BioCAT beamline 18ID of the Advance Photon Source in Chicago is well suited for collecting such data, especially for large molecules for which very low Q data are desired. The research presented in this poster demonstrates how solution structures derived *ab initio* from small- and wide-angle X-ray scattering (SAXS/WAXS) data complement the microscopy approaches (comparable cryo-electron microscopy data will be presented separately by Sacha de Carlo and Eva Nogales).

We have collected scattering data for several proteins or protein fragments of DctD, NtrC, NtrC1 and PspF proteins, four such AAA+ ATPases. Solution structures determined from the scattering data give us insight into regulation and function of these molecular motors. In one case, a regulatory domain adopts two homo-dimeric forms, alternately repressing or derepressing motor assembly by adjacent ATPase domains; in another case, regulatory and ATPase domains cooperate to stabilize the assembled motor. Structures of ATPase in the presence of nucleotide analogs promise to reveal subdomain reorientations that are coupled with conformational changes in the 'second region of homology' and pore region of the ring shaped motors to mediate interaction with the target protein, σ54. Scattering data also yield preliminary models to explain how σ54 binds tightly to the activator in the transition state for ATP hydrolysis.

Use of the Advanced Photon Source was supported by the Department of Energy under contract W-31-109-ENG-38. BioCAT is a Research Center supported by National Institutes of Health, grant RR08630.

Cell-Free Protein Synthesis for High-Through-Put Proteomics
MacConnell Research Corporporation

Evan Dushman, Randal Sivila, and Jennifer Holmes and William P. MacConnell

The production of proteins from cloned DNA sequences is an important process for functional genomic studies and structural analysis, as well as many research applications including pharmaceutical drug discovery. We are developing new methodology and products for cell-free protein synthesis that allow production of up to 100 milligrams of protein using an inexpensive and highly stable wheat germ cell-free system. This system offers a tremendous advantage in simplicity and cost over *in vivo* protein expression methods in that it simplifies or eliminates: vector construction, cell transfection, and uncertainties of host cell synthesis. The method also allows the expression of multiple proteins in parallel, and can begin with PCR-generated DNA templates.

The overall objective of this work is to develop affordable *in vitro* protein synthesis reagents that will produce up to 100 milligrams of highly active protein using a universally applicable protocol. A simple processing instrument is also being developed to automate the protein synthesis reaction steps.

Results thus far demonstrate that our enhanced S30 wheat germ lysate can generate up to 30 milligrams of enzymatically active protein in one reaction. The process begins with double-stranded template DNA that is transcribed into mRNA (non-capped) using T7 RNA polymerase. A typical mRNA transcript is designed to contain the coding domain of the desired protein downstream from a strong ribosome binding sequence such as the TMV 5'UTR. The synthesized protein can be produced from either plasmid or PCR generated template DNA. In the case of PCR template, we begin with genomic DNA that was amplified with a first set of primers, then re-amplified with subsequent primer sets to add the T7, UTR and/or affinity tag sequences to the message or protein.

We have used the system to synthesize seven different proteins of varying sizes from bacterial and mammalian origin. Two of these proteins were successfully purified from the reaction mixture using a 6-his tag affinity purification method. The wheat germ system has been shown to be scaleable in trials where the energy generating reagents were added sequentially or when these components were diffused into the reaction through a permeable membrane. The in vitro method allows for the expression of toxic proteins that are impossible to produce by cellular expression methods. The system can also be used to generate protein from a predicted, but unknown, coding sequence or multiple variations of the same protein.

Several products will arise from this technology that can be sold directly by our company to laboratories throughout the world that perform genomic and proteomic research. We estimate that purified protein can be synthesized by this system for $13 per milligram. These products will save time and labor, improve the outcome of experiments and reduce the cost of small-scale protein production.

DOE Grant Number: DE-FG02-03ER83706