# Genomics:GTL Program Projects

## Harvard Medical School

# Microbial Ecology, Proteogenomics, and Computational Optima

# 1

## Flux Balance Based Whole-Cell Modeling of the Marine Cyanobacterium *Prochlorococcus*

**George M. Church**[1] (g1m1c1@arep.med.harvard.edu), Daniel Segre[1], Xiaoxia Lin[1], Kyriacos Leptos[1], Jeremy Zucker[2], Aaron Brandes[2], Dat Nguyen[1], and Jay MacPhee[1]

Department of Genetics, Harvard Medical School, Boston, MA and [2]Dana-Farber Cancer Institute, Boston, MA

http://arep.med.harvard.edu/DOEGTL/

The marine unicellular cyanobacterium *Prochlorococcus* is the dominant oxygenic phototroph in the tropical and subtropical oceans, and contributes to a significant fraction of the global photosynthesis (Rocap et al, 2003). Our goal in this project is to develop whole-cell mathematical models for studying the metabolism of this cyanobacterium using flux balance based approaches, which has proven very successful in performing whole-cell modeling for a variety of microorganisms (Price et al, 2003). An especially interesting challenge is the inclusion of photosynthesis pathway in our model. Night-day cycles are known to play a central role in the metabolism of *Prochlrococcus*, and different strains are adapted to different light intensities and wavelengths. Flux balance models give the opportunity to study quantitatively the influence of photon fluxes on global cell behavior.

The completion of the *Prochlorococcus* genome sequencing has provided us a promising starting point for building whole-cell flux balance models of this bacterium. By utilizing an automatic bioinformatics pipeline which was recently developed (Segre et al, 2003), we have combined the genome annotation of *Prochlorococcus* MED4, a high-light-adapted strain, with an extensive pathway/genome database, MetaCyc (Karp et al, 2002), and generated a *Prochlorococcus* MED4 pathway database. This organism-specific pathway database is then used to generate flux balance models in which given the stoichiometric matrix representing the metabolic networks and limits on nutrient uptakes, linear programming (LP) or other optimization techniques are used to calculate the flux distribution that reflects the metabolic state of the cell. Our preliminary studies have shown that a substantial number of the biomass components can not be produced with the current identified metabolic networks. This is mainly due to i) incomplete annotation of the genome, for example, not identifying a gene encoding the enzyme catalyzing a metabolic reaction in the biosynthesis pathway of a certain amino acid; and ii) incomplete inclusion of pathways from the

MetaCyc database. In order to generate flux balance models that can capture the primary components of the metabolic networks of *Prochlorococcus* and then can be used to study its genotype-metabolic phenotype relationship under varying conditions, we are currently improving and refining the models by i) using network debugging methods to identify missing reactions/pathways in the constructed *in silico* metabolic network; ii) including additional reactions/pathways based on information from a variety of other sources, such as identification of enzymes through manual search of homologs, proteomic data, existing knowledge about the bacterium's metabolism, etc.

Another important requirement for the construction of whole-cell flux balance models of *Prochlorococcus* is to incorporate an appropriate set of transport reactions, which are currently lacking in the MetaCyc database. Approximately 50 transport proteins have been classified according to the Transport Classification system, which includes substrate specificity, through a combination of TC-BLAST, pfam, COG, and phylogenetic tree analysis (available at http://membranetransport.org). The transport reactions associated with these proteins can be deduced directly from their Transport classification number. We are working closely with the curators of MetaCyc and the Membrane transport database to incorporate these reactions into the pathway/genome database for *Prochlorococcus*.

Upon the successful construction of whole-cell flux balance models for *Prochlorococcus*, we plan to i) investigate how the metabolic network of this cyanobacterium works to enable it grow/live under its natural environmental conditions, in specific, in the light and in the dark; ii) investigate the differences between high-light-adapted strains, for example, MED4, and low-light-adapted strains, for example, MIT9313, by comparing the structures of their metabolic networks and the calculated flux distributions under varying conditions; and iii) investigate the effect of gene knockouts on cellular properties, such as growth rate and photosynthesis, using the MOMA approach developed earlier in the Church lab (Segre et al, 2002). Hypotheses generated with flux balance models will be tested experimentally using expression and proteomic data.

**Reference**


1. Karp PD, Riley M, Paley S, and Pellegrini-Toole A (2002) The MetaCyc Database. *Nucleic Acids Research* **30**(1):59-61.

2. Price ND, Papin JA, Schilling CH, and Palsson BO (2003) Genome-scale microbial in silico models: the constraints-based approach. *Trends Biotechnol*. **21**(4): 162-169.

3. Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, Johnson ZI, Land M, Lindell D, Post AF, Regala W, Shah M, Shaw SL, Steglich C, Sullivan MB, Ting CS, Tolonen A, Webb EA, Zinser ER, and Chisholm SW (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**(6952):1042-1047.

4. Segre D, Vitkup D, and Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc. Nat. Acad. Sci USA* **99**: 15112-7.

5. Segre D, Zucker J, Katz J, Lin X, D'haeseleer P, Rindone W, Karchenko P, Nguyen D, Wright M, and Church GM (2003) From annotated genomes to metabolic flux models and kinetic parameter fitting. *Omics* 7:301-16.

## Lawrence Berkeley National Laboratory

# Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria

# 2

## 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[4], Katherine Huang[1], Kevin Keck[1], Frank Olken[1], Vijaya Natarajan[1], Morgan Price[1], and Yue Wang[2]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]University of California, Berkeley, CA; [3]Howard Hughes Medical Institute, Chevy Chase, MD; and [4]Research Institute for the Genetics and Selection of Industrial Microorganisms, Moscow, Russa

The primary roles of the Computational Core are to curate, analyze, and ultimately build models of the data generated by the Functional Genomics and Applied Environmental Microbiology Core groups. The near-term focus of the computational group has been to build the scientific and technical infrastructure necessary to carry out these roles. In particular, the efforts of the computational group have been directed toward three objectives: genomics and comparative genomics, curation and analysis of experimental data from the other core groups, and modeling. Central to each of these goals has been the development of a comprehensive relational database that integrates genomic data and analyses together with data obtained from experiment.

**VIMSS DB.** At present, well over 100 microbial genomes have been sequenced, and hundreds more are currently in the pipeline. Despite this fact, tools to explore this wealth of information have focused on individual genome sequences. The VIMSS Comparative Genomics database and web-based tools are designed to facilitate cross-species comparison, as well as to integrate experimental data sets with genome-scale functional annotations such as operon and regulon predictions, metabolic maps, and gene annotations according to the Gene Ontology. Over 130 complete genome sequences are represented in the VIMSS Comparative Genomics Database, which is implemented as a MySQL relational database, a Perl library for accessing the database, and a user-friendly website designed for laboratory biologists (http://escalante.lbl.gov). This database is currently being augmented with a novel graph for the efficient query of biological pathways and supporting data. A generic java-based tool for the graphical construction of queries on representations of relational database schema (particular for pathways) in nearly finished and will be applied to VIMSS DB in first quarter 2004.

**Web-Based Tools.** The VIMSS Comparative Genome Browser allows users to align any number of genomes and identifies predicted orthology relationships between genes. Users can save genes of interest for use in the VIMSS Bioinformatics Workbench (VBW), explore individual genes in depth for information about sequence domains, BLAST alignments, predicted operon structure and functionally related genes inferred from a combination of comparative genomics

methods and microarray experiments. The VertiGO comparative gene ontology browser allows users to simultaneously view the genetic complement of any number of genomes according to the Gene Ontology hierarchy. A metabolism browser based on the KEGG metabolic maps allows browsing either the set of enzymes predicted to be present in a single genome, or a comparison highlighting the metabolic differences between two genomes. VBW allows users to create and save lists of genes of interest, and use these lists to investigate phylogenetic relationships by making multiple sequence alignments and phylogenetic trees, as well as apply DNA motif-finding software to identify potential regulatory elements in upstream sequences. Novel motif finding algorithms exploiting the comparative analysis of orthologous proteins have already been accurately difficult motifs such as those from the merR family of regulators of heavy-metal resistance.

**Genome Annotation.** One of the stated goals of the GTL program is to produce next-generation annotation of target genomes including automated gene functional annotations and prediction of gene regulatory features along with validation of these in silico methods. The most fundamental unit of gene regulation in bacteria is the operon, which is a set of genes that are cotranscribed on a single RNA transcript. Because few operons have been characterized experimentally outside the model organisms *E. coli* and *B. subtilis*, in silico operon prediction methods have been validated only in these two organisms. We have therefore made accurate and unbiased operon predictions in all bacteria a priority for the computational group. To avoid bias that might arise from using experimental data from only two organisms, we have opted to avoid the use of experimental data entirely using techniques from the field of unsupervised machine learning, and we used gene expression data to estimate the accuracy of our predictions. Key to the success of this approach has been integrating experimental data from the Functional Genomic Core group into our Comparative Genomics Database to validate our in silico procedures. Using our operon prediction tool, we have established that, contrary to reports in the literature, the bacterium *Helicobacter pylori* has a large number of operons. In addition, by examining unusually large non-coding regions within highly conserved operons, we have identified putative pseudogenes in *Bacillus anthracis* that allow us to make phenotypic predictions about the motility of the sequenced Ames strain. As a critical test of our automated genome annotations, we are hosting a genome annotation jamboree in April at the Joint Genome Institute, in which our automated predictions will be verified by human curators. We expect that our annotations, along with confidence levels, will reduce the manual curation workload allowing participants to focus most of their efforts on scientific hypothesis testing.

**Functional Genomics.** The Functional Genomics Core group is beginning to produce large data sets detailing the response of our target organisms to a variety of stress conditions. The Computational Core group is charged with the responsibility to: store and redistribute these data; assist in the statistical analysis and processing of raw data; and to facilitate comparison of experiments performed with different experimental techniques, different conditions, or different target organisms. As a test case, we have focused most of our efforts in this direction toward gene expression microarray experiments. Among the challenges in the representation of microarray data is developing a data schema that includes both raw and processed data, metadata describing the experimental conditions, and a technical description mapping, for example, each array spot to a corresponding region of the genome sequence and to the set of annotated genes (and their orthologs in other species). We are actively following the development of standards for the representation of this type of data (see Data Management below), and in the meantime have implemented our own simple formats aimed at quick integration with our Comparative Genomics Database. To interpret the results of these experiments, it was necessary

to develop a standard set of procedures for data normalization and significance testing and apply it uniformly to raw data from each experiment set, as processed data from different labs commonly involve slightly different analytical techniques. By establishing common methodologies, and a common repository for different experimental results, we were able to meet the goal of facilitating comparative studies as well as using the functional genomic data to test hypotheses generated from our comparative genomic analysis. The methods have been applied to the analysis of pH, salt and heat stress data from *Shewanella oneidensis*. Results from this analysis will be described.

**Data Management.** During the first year of the project, laboratories in the project began putting in place experimental procedures and are now beginning to produce substantial amounts of data. There is a critical need to define what descriptions of data and experimental procedures (protocols) and factors need to be developed and captured, and to put in place procedures for documenting and recording that information. Recognizing this need, we are in the process of reviewing how experimental procedures are being documented and how experimental factors are being recorded by LBNL affiliated laboratories. This information will be used not only to facilitate information and data acquisition procedures, but also to to enhance and upgrade the BioFiles system for data uploading and the underlying database management system. Working with a consortium of researchers from the wider GTL community we have produced a report on the current status of National Data standards and their advantages and deficiencies and produced a plan for developing standardization of metadata and data representation.

# 3

## Managing the GTL Project at Lawrence Berkeley National Laboratory

**Nancy A. Slater** (naslater@lbl.gov)

Lawrence Berkeley National Laboratory, Berkeley, CA

The effective management of the GTL systems biology project at Lawrence Berkeley National Laboratory (LBNL) is essential to the success of the project. The comprehensive management plan for the project includes milestone planning and project integration, a plan for communicating and collaborating with the project stakeholders, financial management and website updates. In addition, the management plan incorporates reviews by committees, including a monthly Executive Committee review comprised of LBNL leadership, an annual Scientific Advisory Committee review, a biannual Technical Advisory Panel review to ensure that the project's technical development is aligned with related DOE efforts, and a monthly Steering Committee conference call where the project leaders discuss the project's progress and status.

A key responsibility in the project management process is troubleshooting problems related to the scientific and financial management of the project. There is a delicate balance between having adequate resources to achieve the scientific objectives of the project and working within the funding levels of the project. If an area is falling behind on achieving their scientific milestones, the project manager must work closely with the researchers to resolve problems as efficiently and effectively as possible.

**Milestone Planning and Project Integration**

A detailed list of project deliverables and milestones is updated by the PIs at the beginning of each fiscal year. The process of updating and reviewing milestones ensures that the goals of each PI are aligned with the overall goals of the project. These milestones are the basis for an integrated project schedule, which is managed using Microsoft® Project. The project schedule is updated monthly, and progress is reported through progress reports and teleconferences with the PIs. The updated project schedule is posted to the project website, so that all of the collaborators have access to the most recent status of the project.

The project is divided into three separate Core groups, and the integration plan for the project assures that the Core groups work together toward the objectives of the project. The Core Research group leaders are responsible for ensuring smooth operation of their section of the project as well as cooperation with the other groups. For example, the Applied Environmental Microbiology leader is responsible for ensuring that cell culture protocols are acceptable to the Functional Genomics Core, who will ultimately use the cell cultures for experiments. The Functional Genomics Core leader is responsible for ensuring quality control for data production and timely data uploads into the database. The Computational Core leader is responsible for ensuring that data entry, querying, and curation interfaces serve the needs of the other groups, and that the models are useable to biologists outside of the modeling group. The success of each group is interdependent on a well-integrated project team.

**Communication and Collaboration**

The GTL project at LBNL is a collaborative effort between seven institutions, thirteen researchers and their associated laboratories. The project's communications plan consists of a variety of media, including a project website, monthly group meetings, conference calls, an annual retreat, workshops at conferences, and monthly progress reports.

The monthly group meetings include a presentation from one of the Core Research groups, and it is attended by the local, northern California GTL project team members. There are a several conference calls that are held on a regular basis, including a monthly Steering Committee meeting in which all of the researchers participate, a monthly BioFiles conference call in which a representative from each laboratory discusses data generation, uploads and handling, and a quarterly conference call with DOE. The LBNL project has an annual retreat in which the researchers present data and findings related to their area of focus and other laboratory team members (Computer Science Engineers, Microbiologists, Database Managers, Graduate Students, Post Docs, etc.) present posters in a poster forum. The annual retreat has proven to be very successful in building working relationships among the dispersed group. The LBNL GTL project will be participating in several workshops at international conferences in 2004. The monthly progress reports are comprised of input from each researcher, and include updates regarding the status of the milestones, planned work and problems/issues that they encountered.

**Website Updates**

The GTL project at LBNL is the inaugural project for the Virtual Institute of Microbial Stress and Survival (VIMSS), and details regarding the project are located on the world wide web at http://vimss.lbl.gov. This website serves as a tool for communicating the status of the project as well as:

- an overview of the GTL project at Berkeley and links to the key personnel working on the project
- a link to Comparative Genomics Tools such as the Comparative Genome Database, the Genome Browser, and Operon and Regulon Prediction tools
- a link to the BioFiles repository of project data
- a discussion board for project team members to interact and post protocols, questions and solutions
- a job board with available GTL-related positions
- a calendar of upcoming meetings and events

**Financial Management**

Each of the researchers provides input into the annual spend plan for the project. The finances of the project are tracked on a continuous basis, and the researchers receive monthly reports showing actual costs verses the spend plan. The finances of the project are maintained using software packages at LBNL as well as spreadsheets and charts. These tools allow the Project Manager to identify spending trends, so that appropriate can be taken to keep the project aligned with the annual spend plan. The Executive Committee reviews the project financial reports monthly.

# 4

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

**Terry C. Hazen***[1] (TCHazen@lbl.gov), **Hoi-Ying Holman**[1], Sharon E. Borglin[1], Dominque Joyner[1], Rick Huang[1], Jenny Lin[1], **David Stahl**[2], Sergey M. Stolyar[2], **Matthew Fields**[3], **Dorothea Thompson**[3], **Jizhong Zhou**[3], **Judy Wall**[4], H.-C. Yen[4], and **Martin Keller**[5]

*Presenting author

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

**Field Studies**

**Sulfate-reducing bacteria.:** Sediment samples from different depths at the NABIR Field Research Center in the background, Areas 1, 2, and 3 sites have been used for the enrichment of sulfate-reducing microorganisms. Sulfate-reducing enrichments have been positive for sediments in Areas 1 and 2 when lactate or acetate were used as electron donors, and some of the enrichments differ in the capacity to reduce cobalt, chromium, and uranium. Groundwater enrichments from Areas 1, 2, and 3 all displayed sulfate-reduction with different electron donors (lactate, butyrate, acetate, pyruvate) and these enrichments could also reduce iron, cobalt, and chromium. Subsurface sediments from the wells FWB-107 (13.2 m) and FWB-109 (15.4 m) in Area 3 were serially diluted in a basal salts medium that contained lactate and ethanol with different electron acceptors. The results suggested that in the sampled sediments (13 to 15 m) nitrate-reducers were approximately 3500 to 5400 cells/g, iron-reducers 50 to 1700 cells/g, and sulfate-reducers 240 to 1100 cells/g. The predominant population (25%) of the 10-2 sulfate-reducing dilution had 88%

sequence identity with *Desulfosporosinus blif*. Subpopulations that had 95% to 97% sequence identity with *Desulfosporosinus orientis* constituted for an additional 37% of the library. Other clones had 98% sequence identity with *Clostridium chromoreductans*.

**Clone libraries**. Since stress response pathways are clustered on chromosomal DNA fragments and generally vary in length from 20-40 kb, it is essential to clone large DNA fragments to capture entire pathways. We have developed effective DNA extraction methods and vector/host systems that allow stable propagation of large DNA fragments in *E. coli*. Processed environmental samples are embedded in agarose noodles for protein digestion and release of high molecular weight DNA. In stressed environments, organism concentrations are often very low, so we have developed a method for increasing the concentration of large DNA by amplification with a phage polymerase. After amplification, the DNA is partially digested with restriction enzymes, and size-selected by agarose gel electrophoresis. It is then ligated to fosmid arms and packaged into phage lambda particles that are used to infect *E. coli*. The microbial diversity of the libraries is determined with Terminal Restriction Fragment Polymorphism (T-RFLP). Large fragment DNA has been extracted and amplified from 15 NABIR FRC samples (comprising 3 areas at various depths). Small insert DNA libraries have been constructed from most of these samples, and large insert DNA libraries are in various stages of construction. T-RFLP and DNA sequencing are being used to quality control the resulting libraries.

**Enrichments**. Seven *Desulfovibrio* strains were isolated from lactate-sulfate enrichment of sediment taken from the most contaminated region of Lake DePue, IL. Their 16S rRNA and dsrAB genes were amplified and sequenced. They all were identical to each other and virtually identical to the corresponding genes from *D. vulgaris* Hildenborough. One mismatch was observed in the16S rRNA gene and one in dsrAB. Different fragment patterns confirmed that the DePue isolates were similar but not identical to *D. vulgaris* Hildenborough. Pulse field electrophoretic analysis of I-CeuI digests revealed that both isolates had five rRNA clusters, the same as *D. vulgaris* Hildenborough. However, the length of one chromosomal segment in the DP isolates was considerably shorter than the corresponding fragment from *D. vulgaris* Hildenborough, suggesting the presence of a large deletion in the genomes of the isolates (or insertion in *D. vulgaris* Hildenborough).

**Culture and Biomass Production**

**Defined Media – Growth**. A defined medium for optimal growth and maximum reproducibility of *Desulfovibrio vulgaris* was developed for biomass production for stress response studies. The medium was optimized by evaluating a variety of chemical components, including the removal of yeast extract, excess sulfate, and Fe, and redox conditions to optimize cell density and generation times, and to reduce lag times. Growth was monitored using direct cell counts, optical density, and protein concentration. The generation time for *D. vulgaris* in the original Baar's medium was 3 h, reaching a maximum density of $10^8$ cells/ml and 0.4 $OD_{600\ nm}$. The generation time for *D. vulgaris* on LS4D was 5 h, with a maximum cell density of $10^9$ cells/ml and a 0.9-1.0 $OD_{600\ nm}$. LS4D is well suited for the monitoring protocols, as well as the equipment and large scale processing needed for biomass production.

**Dual culture systems**. Co-cultures of two different *Desulfovibrio* species (*Desulfovibrio vulgaris* Hildenborough and *Desulfovibrio* sp.PT2) syntrophically coupled to a hydrogenotrophic methanogen (*Methanococcus maripaludis*) on a lactate medium without sulfate has been established and characterized. No appreciable

growth was observed in 50 mM lactate for single-organism cultures. Following optimization of the ionic composition (MgCl2 and NaCl) of the medium, stable co-cultures were established having generation times of 25h-1 and 35 h-1 for *D. vulgaris* and *Desulfovibrio* sp. PT2 co-cultures respectively. Both co-cultures degraded lactate to acetate, methane, and carbon dioxide. No other organic acids were detected during the course of experiments. Approximately 1mol of acetate and 1mol of methane was produced from two mole of lactate by both co-cultures during most active period of growth. The stability of established methanogen-SRBs co-cultures (*Desulfovibrio vulgaris* or *Desulfovibrio* sp. PT2 with *M. maripulidis*) was confirmed by serial transfer (six times).

**Biofilm reactors**. Initial characterization of *Desulfovibrio vulgaris* growth as a biofilm was evaluated using a 600ml biofilm reactor containing 3mm glass beads as growth substratum and the B3 culture medium (16mM lactate and 28 mM sulfate). The ratio of flow rates through an internal recirculation loop to influent was maintained at 100:1, evaluating two different influent flow rates (0.5ml/min or 30ml/hr). Formation of a loose biofilm was associated with significant gas accumulation within the reactor. The system in now being modified to incorporate a gas trap in the re-circulation loop.

**FairMenTec (FMT) chemostat**. A pilot run with *Desulfovibrio vulgaris* Hildenborough in the FMT bioreactor in chemostat mode was completed. The bioreactor was operated using the LS4D medium with 45mM lactate, 50 mM sulfate, and Ti-citrate at 1/3 standard formulation (subsequent batch cultures have shown improved growth with further reduction of the Ti-citrate to 1/6 standard formulation). Varying flow rates and medium compositions were evaluated.

**Oxygen Stress Experiments**

**Protocols**. Since episodic exposure to air or oxygenated ground water is common at contaminated sites, we decided to focus on oxygen stress of *D. vulgaris* for our initial studies. To accommodate the all the investigations that would require simultaneous harvesting of biomass for studies on proteomics, transcriptomics, metabolomics and phenotypic studies a batch culture system was developed for 2000 ml cultures that could be sparged with nitrogen or air to control stress in water baths using rigorous quality control on culture age, sampling, defined media, chain of custody, and harvesting times and techniques.

**Phenotypic responses**. Desulfovibrio vulgaris enters a new phenotypic state when confronted with a sudden influx of oxygen. Using SEM and TEM microscopy we observed that during the first 24-72 h of exposure to air *D. vulgaris* cells are negatively aerotactic, gradually they loose their flagella, and begin to elongate, by 20 days exposure they are 3-4 times larger and have a well developed exopolysaccharide sheath. At all times the cells were viable and recovered when put back under anaerobic conditions. Real-time analysis using Synchrotron Fourier Transform Infrared Spectromicroscopy enabled us to determine quantitative changes in peptides and saccharides in the living cells during exposure to air, thus providing the exact timing of cell changes in the stress response. During the early phase of the exposure, we observed decreases in total cellular proteins as well as changes in the secondary structures of proteins that are indicative of the changing of the local hydrogen-bonding environments and the presence of granular protein. During the late phase of the exposure, we observed the production of polysaccharides, concomitant with the production of the external sheath. The S-FTIR also demonstrated that the cells were viable within the sheath at 20 days exposure. Phospholipid fatty acid (PLFA) analysis confirmed that no biomass was loss during air sparging of stationary phase cells.

In addition, no change in the PLFA patterns were observed during air sparge, indicating neither cell growth nor death occurred. The PLFA extraction is being developed as a method for routine monitoring of cultures during biomass production and stress studies. Databases of lipid signatures of *D. vulgaris* during various growth conditions are being developed to augment the information produced from other VIMSS collaborators on proteomics and functional genomics.

# 5

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

**Jay Keasling***[1] (keasling@socrates.berkeley.edu), 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], Vincent Martin[1], Aindrila Mukhopadhyay[1], 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), and **Jizhong Zhou**[4] (zhouj@ornl.gov)

*Presenting author

[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 Corporation, San Diego, CA

**Introduction:** 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.

**Transcript analysis:** *D. vulgaris* is a δ-Proteobacteria with a genome size of approximately 3.6 Mb. Whole-genome microarrays of *D. vulgaris* were constructed using 70-mer oligonucleotides. All ORFs in the genome are represented with 3,471 (97.1%) unique probes and 103 (2.9%) non-specific probes that may have cross-hybridization with other ORFs. The microarrays were employed to investigate the global gene expression profiles of *D. vulgaris* in response to elevated salt and nitrite concentrations as well as exposure to oxygen. Approximately 370 ORFs were up-regulated (≥ 3-fold) and 140 ORFs were down-regulated when *D. vulgaris* cells were treated with 0.5 M NaCl for 0.5 hour. For example, genes involved in glycine, betaine, or proline transport were up-regulated 5-, 19- and 26-fold, respectively. Almost half of those genes with significant changes in expression are predicted as conserved hypothetical or hypothetical proteins. After 4-hour treatment, approximately 140 ORFs were up-regulated and more than 700 ORFs were down-regulated. Patterns of gene expression were distinctly different between time points. With 1 mM nitrite, *D. vulgaris* exhibited a lag phase of 28 h compared to a 5 h lag

phase in controls without nitrite addition. Strong nitrite treatment (5 or 10 mM) triggered a transient growth arrest and growth resumed gradually after 5 hours, suggesting the ability of *D. vulgaris* to overcome the toxicity of nitrite. Transcriptional profiling analysis was carried out following nitrite (10 mM) treatment. Transcripts highly up-regulated throughout the 5 h following nitrite shock included genes encoding two iron-sulfur cluster-binding proteins (65- and 15-fold) and a hybrid cluster (Fe/S) protein (24-fold). All three ORFs are annotated as redox-active proteins, and the hybrid cluster protein has been specifically proposed to participate in nitrogen metabolism. Surprisingly, the nitrite reductase genes were only moderately up-regulated (3-fold) as well as the formate dehydrogenase genes.

**Protein analysis:** A combination of Differential In-Gel Electrophoresis (DIGE), Isotope-Coded Affinity Tags (ICAT), and comprehensive proteome analyses were used to investigate the response of the *D. vulgaris* proteome to heat shock and $O_2$ stress. DIGE analysis of heat-shock stress response identified a total of 650 proteins. Sixty-three (63) proteins showed differences between the heat shocked (30 min) and control conditions. Using the complementary ICAT analysis we were able to identify a total of 219 proteins out of the *D. vulgaris* proteome. Out of this pool of proteins, 7 stress related proteins were identified. Similar analysis was also done with $O_2$-stressed cells. Based on cysteine containing tryptic peptides, a total of 92 proteins were identified. Among the identified proteins, 40 showed differences between the $O_2$-stressed and control conditions and of these at least 6 are known to be involved in $O_2$-stress response. Total comprehensive proteome analysis of *D. vulgaris* was also used to investigate differential protein expression induced by $O_2$-stress. Cellular tryptic-digested proteins from control and stressed cultures were analyzed by 3D $\mu$LC-MS-MS. A total of 1,791 unique proteins were identified.

**Protein complex analysis:** Based on the preliminary DIGE analysis of heat shock response in *D. vulgaris*, HSP70 (ORF00281) was identified as being involved in this stress condition. Western analysis using antibodies to the *E. coli* homolog (63% sequence identity) showed enhanced production of ORF00281 (Hsp70). The Anit-HSP70 antibody was then used to study bait-prey interactions in whole cell protein extracts from the heat shock condition using the Co-Immunoprecipitation kit for immobilization. Approximately 7 "pulled down" proteins bands were observed as possibly interacting proteins with HSP70. These bands were gel extracted and further analyzed by LC-MS-MS. To generate tagged proteins for identifying protein complexes in *D. vulgaris*, we have also explored the application of the IBA Strep-tag vector system for generating single chromosomal copies of genes fused to the tag sequence. We have generated a fusion of *dnaK* with the tag and have it integrated into the chromosome of *D. vulgaris* in single copy to determine the effectiveness of this system for providing complexes for proteomics analysis.

**Metabolite analysis:** We have developed a hydrophilic interaction chromatography method coupled to MS/MS detection to separate and identify nucleotides and redox cofactors. In addition, CE-MS methods were developed to analyze a variety of metabolites, including amino acids, nucleic acid bases, nucleosides, nucleotides, organic acid CoAs, redox cofactors, and the metabolic intermediates of glycolysis, the TCA cycle and the pentose phosphate pathway. All the methods were validated using *E. coli* cell extracts. Approximately 100 metabolites can be separated and identified. The development of an efficient method to obtain *D. vulgaris* metabolite extracts and its application to analyze stress responses in *D. vulgaris* are in progress.

**Development of a genetic system:** In efforts to improve the genetic versatility of *D. vulgaris*, spontaneous mutants resistant to either nalidixic acid or rifampicin were selected. These antibiotic resistances will allow counter-selection of sensitive *E. coli*

donors in conjugation experiments. Additional effort has been made to screen antibiotic sensitivity and resistance of *D. vulgaris*. The wild type was sensitive to G418 (400 μg/ml), ampicillin (20-50 μg/ml), carbinicillin (20-50 μg/ml) and resistant to gentamycin. The drug resistance markers present on many routinely used cloning vectors confer resistance to these antibiotics. Marker exchange mutagenesis of a number of regulatory genes is in progress by a procedure that will introduce molecular barcodes into the deletion sites. Sucrose sensitivity will be used to enrich for the second recombination event necessary to delete the wild-type copies of the target genes. Interestingly, we found that sucrose sensitivity is not expressed well in all *Desulfovibrio* strains. To further streamline methods for gene knockout, a vector system that uses a single cross-over event for gene deletion has been created. A 750-bp internal gene sequence flanked by 20 base pair UP and DOWN barcodes will used to simultaneously knock out and barcode each gene. Conjugal transfer using *E. coli* will be used to transform *D. vulgaris* with the suicide knockout vectors. The single cross-over gene deletion system also attempts to address issues of polar mutations. Additionally, a *lacZ* reporter will be incorporated into the site of gene deletion. Methylumbelliferyl β-D-galatoside, a fluorescent substrate the β-galactosidase reporter will be used for colony screening under anaerobic conditions. Finally, experiments to generate a library of transposon mutants are also underway. Putative mutants have been generated and will be screened for the presence and copy number of the transposon, stability of the antibiotic resistance, and randomness of the insertion.

## Oak Ridge National Laboratory and Pacific Northwest National Laboratory

# Genomics:GTL Center for Molecular and Cellular Systems

*A Research Program for Identification and Characterization of Protein Complexes*

# 6

## Establishment of Protocols for the High Throughput Analysis of Protein Complexes at the Center for Molecular and Cellular Systems

**Michelle V. Buchanan**[1] (buchananmv@ornl.gov), Gordon Anderson[2], Robert L. Hettich[1], Brian Hooker[2], Gregory B. Hurst[1], Steve J. Kennel[1], Vladimir Kery[2], Frank Larimer[1], George Michaels[2], Dale A. Pelletier[1], Manesh B. Shah[1], Robert Siegel[2], Thomas Squier[2], and H. Steven Wiley[2]

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

The first year of the Center for Molecular and Cellular Systems focused on evaluating methods for the efficient identification and characterization of protein complexes, identifying "bottlenecks" in the isolation and analysis processes, and developing approaches that could eliminate these bottlenecks. Oak Ridge National Laboratory (ORNL) and Pacific Northwest National Laboratory (PNNL) staff worked closely together to develop an integrated process for protein complex analysis. Emphasis has been placed on developing robust protocols that are adaptable to high throughput isolation and analysis methods. Progress has been made in all five major program areas—molecular biology, organism growth standardization, protein complex isolation/ purification, protein complex analysis, and bioinformatics/computation. During this first year, we have evaluated a two-phased approach to identify protein complexes. The first is an exogenous bait approach using one or more purified proteins to pull down the components of the associated protein complex. The second is an endogenous approach involving the *in vivo* expression of tagged proteins that are used to pull down the components of the associated protein complex. These complementary approaches each have their advantages. The first permits the high-throughput isolation of complexes from a single sample grown under defined conditions, while the latter permits the identification of complexes under cellular conditions, plus it can be combined with the development of new imaging methods to identify synthesis, turnover, and complex localization in real time. To test the established protocols two organisms were employed, *Rhodopseudomonas palustris* and *Shewanella oneidensis*. Techniques were optimized and standard protocols were established for endogenous complex isolation and exogenous complex isolation that will be deployed in year two of the project.

Considerable progress has also been made in advancing capabilities for the characterization of protein complexes that will minimize current bottlenecks, reduce the amount of sample required, and automate sample handling and processing. We have

made progress toward using an affinity-labeled crosslinker that allows selective isolation and subsequent mass spectrometric analysis of crosslinked peptides. Microfluidic technologies that reduce the amount of sample required for analysis and decrease the time required for separation have been applied to the analysis of peptides from protein complexes. Automated trypsinization and sample processing protocols have been developed that are designed around a 96-well format. Imaging of microbial cells, based upon introduction of fluorescent labels onto target proteins, has also been pursued. Automation of key parts of the cloning and complex isolation pipeline was initiated. Particular emphasis was given in this first year in establishing a common laboratory information management system (LIMS) and sample-tracking system that would facilitate distributed workflow across multiple laboratories. Results from the first year of this project have led to the design of a single, high-throughput production pipeline that will integrate efforts at both ORNL and PNNL. This will allow the high throughput analysis of hundreds of complexes during the next year. This pipeline will use complementary pull down methods, both endogenous and exogenous methods, to isolate protein complexes and provide greater confidence in complex characterization. This pipeline will be flexible to allow improved technologies to be incorporated as they are developed.

*7*

## Isolation and Characterization of Protein Complexes from *Shewanella oneidensis* and *Rhodopseudomonas palustris*

**Brian S. Hooker**[1] (Brian.Hooker@pnl.gov), Robert L. Hettich[2], Gregory B. Hurst[2], Stephen J. Kennel[2], Patricia K. Lankford[2], Chiann-Tso Lin[1], Lye Meng Markillie[1], M. Uljana Mayer-Clumbridge[1], Dale A. Pelletier[2], Liang Shi[1], Thomas C. Squier[1], Michael B. Strader[2], and Nathan C. VerBerkmoes[2]

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

As part of the Center for Molecular and Cellular Systems pilot project, we have been evaluating both endogenous and exogenous approaches for the robust isolation and identification of protein complexes. Exogenous isolation uses bait proteins to capture the protein complexes. To evaluate various exogenous isolation approaches, five complexes with differing physical characteristics were employed, both stable and transiently associating protein complexes. These complexes included RNA polymerase, the degradosome, and oxidoreductase, all stable protein complexes of varying complexity, and protein tyrosine phosphatase (Ptp) and methionine sulfoxide reductase (Msr), which are signaling proteins that form transient protein complexes. Evaluation of several different approaches has shown that covalent immobilization of the affinity reagent to a solid support works well to isolate the protein complex away from nonspecifically bound proteins, whether this involves direct bait attachment or the immobilization of an antibody against the bait or epitope tag. Approaches evaluated include covalent attachment of bait protein to glass beads that were subsequently used to capture protein complexes and expression of bait proteins with 6xhis tags, which were used to isolate complexes with nickel-chelating resins.

For endogenous complex isolation, we have developed a convenient, broad host range plasmid system to prepare tagged proteins in the native host. A series of expression vectors have been developed that can be used to transfect *E. coli* or *R. palustris*. These expression vectors have been constructed based on the broad host

range plasmid pBBR1MCS5. This vector was modified to contain the Gateway® pDEST multiple cloning region that allows site specific recombination cloning of targets from Gateway® entry plasmid. Four modified Gateway® destination vectors were constructed that can be used for expression of 6x histidine (6xhis) or glutathione(GST), N- or C- terminally tagged fusion proteins. Using this approach, methods have been developed to purify complexes using a double affinity approach (TAP) and complexes of suitable amounts and purity have been obtained for mass spectrometry evaluation. We have cloned a total of 22 *R. palustris* genes into these expression vectors to test expression and affinity purification methods for isolation of protein complexes using different affinity tags. The tested genes included those which code for proteins that are components of GroEL, GroES, ATP synthase, $CO_2$ fixation, uptake hydrogenase, ribosome, photosynthesis reaction center, Clp protease, and signal recognition. Results suggest while there was no one affinity tag which worked well for all genes tested, there was at least one fusion protein that expressed well for each targets tested. The 6xhis and V5 tag combination, does in fact yield a highly purified product in the test cases examined to date. We have therefore focused our effort on using this TAP purification protocol, using the pBBRDEST-42 plasmid as it encodes both the V5 and 6xhis tags. This approach has been incorporated as a part of standard protocols in a high throughput system and a panel of 200 *R. palustris* genes are being processed to serve as the pilot group for this automated approach.

As a benchmark for developing and evaluating affinity-based methods for isolating molecular machines, we carried out a conventional biochemical isolation (sucrose density gradient centrifugation) of the *R. palustris* ribosome, followed by both "bottom-up" and "top-down" mass spectrometric analysis of the protein components of this large, abundant complex. We have identified 53 of the 54 predicted protein components of the ribosome using by the "bottom-up" method, and obtained accurate intact masses of 42 ribosomal proteins using the "top-down" approach. Combining results from these two approaches provided information on post-translational modification of the ribosomal proteins, including N-terminal methionine truncation, methylation, and acetylation.

# 8

## Bioinformatics and Computing in the Genomics:GTL Center for Molecular and Cellular Systems - LIMS and Mass Spectrometric Analysis of Proteome Data

**F. W. Larimer**[1] (larimerfw@ornl.gov), G. A. Anderson[2], K. J. Auberry[2], G. R. Kiebel[2], E. S. Mendoza[2], D. D. Schmoyer[1], and M. B. Shah[1]

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

Scientists at the Oak Ridge National Laboratory/Pacific Northwest National Laboratory (ORNL/PNNL) Genomics:GTL (GTL) Center for Molecular and Cellular Systems are generating large quantities of experimental and computational data. We have developed a prototype Laboratory Information Management System (LIMS) for data and sample tracking of laboratory operations and processes in the various laboratories of the Center. We have also developed a mass spectrometry data analysis system for automating the mass spectrometry data capture and storage, and computational proteomic analysis of this data.

**A Laboratory Information Management System for the GTL Center for Molecular and Cellular Systems**. The Laboratory Information Management System (LIMS) for the GTL Center for Molecular and Cellular Systems is a central data repository for all information related to production and analysis of GTL samples. It maintains a detailed pedigree for each GTL sample by capturing processing parameters, protocols, stocks, tests and analytical results for the complete life cycle of the sample. Project and study data are also maintained to define each sample in the context of the research tasks that it supports.

The LIMS system is implemented using the Nautilus™ software from Thermo Electron Corporation. This software provides a comprehensive yet extensible framework for a LIMS that can be customized to meet the requirements of the GTL project. Nautilus uses client/server architecture to access data maintained in a central Oracle database and presents an interface based on the Windows Explorer paradigm. The latest Nautilus release includes Web access and this will be added to GTL LIMS in the near future.

The LIMS is configured by establishing workflows that parallel the processing steps completed in the laboratory. For each process it is necessary to define the laboratory environment (stocks, storage locations, instruments, protocols), identify the items to track, the process parameters to collect, the tests that will be conducted, and the test results that will be reported. This information is then used to develop LIMS workflows that will ensure the collection of all critical data.

The initial GTL LIMS system configuration has been completed. This required customization of Nautilus to include additional GTL data items such as primers, genes, and vectors, and programmatic extensions to do GTL specific tasks such as copying files to the central file server and displaying files stored on the central file server. Program extensions also had to be developed to handle some of the processing steps for stocks stored in 96-well plates.

Future plans for the LIMS include additional reporting capabilities, integration with the mass spec data analysis pipeline, barcode implementation, and refinement of the process workflows.

**PRISM Mass Spectrometry Proteomic Data Analysis System**. The Proteomics Research Information Storage and Management (PRISM) System manages the very large amounts of data generated by the mass spectroscopy facility and automatically performs the automated analytical processing that converts it into information about proteins that were observed in biological samples. PRISM also collects and maintains information about the biological samples and the laboratory protocols and procedures that were used to prepare them.

PRISM is composed of distributed software components that operate cooperatively on a network of commercially available PC computer systems. It uses several relational databases to hold information and a set of autonomous programs that interact with these databases to perform much of the automated file handling and information processing. A large and readily expandable data file storage space is provided by a set of storage servers. The basic database software is a commercial product, but the database schemata and content and the autonomous programs have all been developed in-house to meet the unique and continually evolving requirements of the MS facility.

PRISM has been in continuous operation since March 2000, and has been continually upgraded. There have been four major upgrade cycles, and numerous minor ones, including the addition of new functionality and the expansion of capacity as new instruments are added to the facility. Most recently, PRISM has been upgraded

to maintain inter-system tracking information for GTL samples and the ability to maintain and process them in their as-delivered format (96-well plates).

PRISM manages data and research results for all of the mass spec based proteomics studies in our laboratory; this includes over 100 research campaigns or lines of investigation. This research has resulted in 15334 datasets from a number of different mass spectrometers. These datasets have required 41491 separate analysis operations to extract peptide and protein identifications. The total raw data volume managed by PRISM is in excess of 15 Tera bytes. The current rate of production results in approximately 800 datasets per month with significant increases expected in FY04.

**Data Abstraction Layer (DAL).** The DAL is middleware that will provide a level of abstraction for any data storage system in the proteomics pipeline (LIMS, Freezer Software, PRISM, etc.). It will provide a generic interface for building tools and applications that require access to the experimental data and analysis results. It will also allow the pipeline data to be extended without making changes in the manner in which an application already looks at the data. For example, it could be used to facilitate a query performed utilizing proteomic data originating from both PNNL and ORNL. The DAL will be used to provide an interface to the pipeline data as required by selected bioinformatics/analysis tools.

# 9

## Advanced Computational Methodologies for Protein Mass Spectral Data Analysis

**Gordon Anderson**[1] (gordon@pnl.gov), Joshua Adkins[1], Andrei Borziak[2], Robert Day[2], Tema Fridman[2], Andrey Gorin[2], Frank Larimer[2], Chandra Narasimhan[2], Jane Razumovskaya[2], Heidi Sophia[1], David Tabb[2], Edward Uberbacher[2], Inna Vokler[2], and Li Wang[2]

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

Completed analysis of a variety of genomes has lead to a revolution in the methods and approaches of what was traditionally protein biochemistry. Now, an analysis of a variety of protein functions can be undertaken on a genome wide level. Among some of the most interesting and complicated functions of proteins is their nature to form higher order functional complexes. Using a combination of protein pull-down techniques and combined capillary liquid chromatography/mass spectrometry (LC/MS) as a sensitive detector for proteins, new protein complexes are being identified as part of the Center for Molecular and Cellular Systems. Computational tools are being developed to assist in the interpretation of these data. For example, complications in these data arise from non-specific and transient protein interactions. Imperfect bioinformatic tools for peptide identifications that lead to protein identifications found in these complex pull-downs is also a problem. We are using the clustering program, OmniViz, as a tool for discovery of protein complexes in this combination of complicating protein identifications. This includes the ability to view various experiments in a virtual 1D dimension gel format to aid biologists in looking at the results and adjustable features that can be used to compare different ratios of sensitivity and specificity in the putative protein complexes. We are automating the process, leading to standardized approaches and reports for protein components of complexes.

Improved scoring algorithms for matching theoretical tandem mass spectra of peptides to observed spectra are being developed to replace existing scoring algorithms such as that used by SEQUEST. The likelihood of matches can be estimated by probabilistic analysis of fragment ions matches rather than computationally expensive cross-correlation. This greatly improves the speed and accuracy of the peptide scoring system. A computational system for peptide charge determination has also been developed with 98% accuracy using statistical and neural network methods. This allows a several-fold speedup in calculation time without loss of information.

Methods for *de novo* sequencing to construct sequence tags from MS/MS data have also been developed using a statistical combination of informational elements including the peaks in the neighborhood of expected B and Y ions. The approach utilizes all informational content of a given MS/MS experimental data set, including peak intensities, weak and noisy peaks, and unusual fragments. The 'Probability Profile Method' is capable of recognizing ion types with good accuracy, making the identification of peptides significantly more reliable. The method requires a training database of previously resolved spectra, which are used to determine "neighborhood patterns" for peak categories that correspond to ion types (N- or C-terminus ions, their dehydrated fragments, etc.). The established patterns are applied to assign probabilities for experimental spectra peaks to fit into these categories. Using this model, a significant portion of peaks in a raw experimental spectrum can be identified with a high confidence. PPM can be used in a number of ways: as a filter for peptide database lookup approach to determine peptides with post-translational modifications or peptide complexes, *de novo* approach and tag determination.

# 10

## High-Throughput Cloning, Expression and Purification of *Rhodopseudomonas palustris* and *Shewanella oneidensis* Affinity Tagged Fusion Proteins for Protein Complex Isolation

**Dale A. Pelletier**[1*] (pelletierda@ornl.gov), Linda Foote[1], Brian S. Hooker[2], Peter Hoyt[1], Stephen J. Kennel[1], Vladimir Kery[2], Chiann-Tso Lin[2], Tse-Yuan Lu[1], Lye Meng Markillie[2], and Liang Shi[2]

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

This poster will describe the approaches and progress in the joint Oak Ridge National Laboratory/Pacific Northwest National Laboratory (ORNL/PNNL) Center for Molecular and Cellular Systems pilot project on protein complexes. We have adopted the following process design for isolation of protein complexes: (1) construct adaptable plasmids for expression in multiple organisms, (2) adapt standard gene primer design, (3) PCR amplify target genes, (4) clone into donor vector/expression vectors, and (5) express in selected organisms for exogenous and endogenous complex isolation.

We have developed software that designs appropriate PCR primers flanking each gene such that any gene can be amplified from genomic DNA. The resulting PCR products can be directly recombined into entry vectors. We have used the Gateway® cloning system (Invitrogen) to produce entry clones that can be recombined into our modified broad host range expression vectors which contain ori genes compatible with replication in a variety of bacterial hosts.

We have performed PCR amplification from host genomic DNA in 96-well format and shown, using generic conditions, that 60-70% of the reactions yield the predicted size products. We have previously demonstrated automated PCR amplification, cleanup and gel analysis using liquid handling robots and are transitioning to high-throughput hardware. We have successfully PCR amplified approximately 80 *R. palustris* genes and cloned 40 into expression vectors. Twenty of these constructs have been electroporated into *R. palustris*. To date high-throughput electroporation has not been implemented but such 96-well systems are commercially available and will be tested. Plans for automation at this step include an automated colony picker and subsequent robot directed plasmid preps for QA and long term cataloging and storage. We have also successfully cloned over 30 *S. oneidensis* genes into expression vectors using the Gateway® system. Over 20 of these constructs have been successfully expressed in both *E. coli* and *S. oneidensis*.

Expression in *E. coli* and in hosts *R. palustris* and *S. oneidensis* has to date been evaluated primarily using manual processes. Cell samples from relatively large cultures are lysed and IMAC or TAP isolations are completed followed by verification of product by SDS-PAGE and/or Western blot. Tagged *S. oneidensis* proteins expressed in *E. coli* and *S. oneidensis* for exogenous bait experiments are then purified using single-step IMAC on a Qiagen Biorobot 3000 LS. Milligram quantities of up to 12 proteins in parallel have been purified using this automated system.

## Sandia National Laboratories

# Carbon Sequestration in *Synechococcus*

*From Molecular Machines to Hierarchical Modeling*

# 11

## Modeling Cellular Response

**Mark D. Rintoul** (rintoul@sandia.gov), Steve Plimpton, Alex Slepoy, and Shawn Means

Sandia National Laboratories, Albuquerque, NM

While much of the fundamental research on prokaryotes is focused on specific molecular mechanisms within the cell, the aggregate cellular response is also important to practical problems of interest. In this poster, we present results for two computational models of cellular response that take spatial effects into consideration. The first model is a discrete particle code where particles diffuse and interact via Monte Carlo rules so that species concentrations track chemical rate equations. This type of model is relevant to cases where there are a small number of interacting particles, and the spatial and temporal fluctuations in particle number can play a significant role in affecting cellular response. Results with this code are shown for a simulation of the carbon sequestration process in *Synechococcus WH810*. The second model utilizes continuum modeling focusing on carbon concentrations inside and outside of the cell, in an effort to understand carbon transport by *Synechococcu*s within a fluid-dynamic marine environment. It is based on solving partial differential equations on a realistic geometry using finite element methods.

# 12

## The *Synechococcus* Encyclopedia

Nagiza F. Samatova[1], **Al Geist**[1] (gst@ornl.gov), Praveen Chandramohan[1], Ramya Krishnamurthy[1], Gong-Xin Yu[1], and Grant Heffelfinger[2]

[1]Oak Ridge National Laboratory, Oak Ridge, TN and [2]Sandia National Laboratories, Albuquerque, NM

*Synechococcus sp*. are abundant marine cyanobacteria known to be important to global carbon fixation. Although the genome sequencing of *Synechococcus sp*. is complete by the DOE JGI[1], the actual biochemical mechanisms of carbon fixation and their genomic basis are poorly understood. This topic is under both experimental and computational investigation by several projects including Dr. Brian Palenik's DOE MCP project, SNL/ORNL GTL Center[2] and others. These projects have been generating heterogeneous data (e.g., sequence, structure, biochemical, physiological and genetic data) distributed across various institutions. Integrative analysis of these data will yield major insights into the carbon sequestration behavior of

*Synechococcus sp*. However, such analysis is largely hampered by the lack of a knowledgebase system that enables an efficient access, management, curation, and computation with these data as well as comparative analysis with other microbial genomes. To fulfill these requirements, the *Synechococcus Encyclopedia* is being created as part of the SNL/ORNL GTL Center.

The completed sequencing of *Synechococcus sp.* has allowed having a reference axis upon which any type of annotation can be layered. Not only can genomic features such as genes and repeats be placed upon such a reference, but it is also possible to map a variety of other data such as operons and regulons, mutations, phenotypes, gene expressions (e.g., microarray, phage display, mass spec, 2-hybrid), pathway models, protein interactions, and structures. A major benefit of such feature mapping is that each of these annotations can be cross-referenced to each other. The *Synechococcus* Encyclopedia takes advantage of this fact to allow users to view and track a variety of biological information associated with the genome and to enable complex queries across multiple data types.

In order to make the exploration and in-depth analysis of genome information easier, one needs appropriate ways to browse and query the corresponding data. The World Wide Web interface of the *Synechococcus* Encyclopedia was built up with these specifications in mind. It offers a number of ways to retrieve information about a genome. For instance, an advanced search capability is available to combine several search criteria and retrieve detailed information about any intricate features. Moreover, the search can be restricted to a genome region of interest, molecular function, biological process, or cellular component.

To ease data retrieval, all output reports are presented in tabular format and maybe conveniently downloaded as tab-delimited text. If desired, the tables can be easily customized, e.g. adding or removing features, or changing the sort order according to several data fields. Data are accessible through a variety of interactive graphical viewers. Furthermore, the retrieved data entries can be further explored by launching complex analysis tools or linking to other data collections such as Swiss-Prot, Pfam, InterPro, PDB, etc.

The *Synechococcus* Encyclopedia comprises information at various levels: genome sequence, structure, regulation, protein interactions, systems biology. For example, at the genome sequence and annotation level, it includes the protein- and RNA-coding genes, Pfam domains, Blocks motifs, InterPro signatures, and COG- and KEGG-based functional assignments. At the structure level, it presents secondary and tertiary structural models predicted by PROSPECT[3] and other tools, SCOP-based functional assignments, and FSSP profiles for homologous protein sequences. At the regulation level, it integrates data about promoters, transcription factors, and pathway models as well as microarray data.

The *Synechococcus* Encyclopedia is accessible at http://www.genomes-to-life.org/syn_wh. The generic data model and data integration, search and retrieval engine that it is based on, makes it possible to set up similar knowledgebases for other bacterial species. New functionalities for multi-genome integration and comparative analysis of genomes are being developed to facilitate better understanding of genomic organization and biological function. Moreover, other features such as annotation and curation services, data provenance, and security will be added in the near future.

**References**

1. http://genome.ornl.gov/microbial/syn_wh

2. Grant Heffelfinger (gsheffe@sandia.gov) and Al Geist (gst@ornl.gov); http://www.genomes2life.org

3. Ying Xu (xyn@ornl.gov) and Dong Xu (xudong@missouri.edu); http://compbio.ornl.gov/structure/prospect/

# 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 (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. Our project includes both an experimental investigation as well as significant computational efforts to develop and prototype new computational biology tools. Several elements of this effort will be discussed including the development of new methods for high-throughput discovery and characterization of protein-protein complexes and novel capabilities for inference of regulatory pathways in microbial genomes across multiple sources of information. Our progress developing new computational systems-biology methods for understanding the carbon fixation behavior of *Synechococcus* at different levels of resolution from the cellular level to ecosystem will also be discussed. More information about our project and partners can be found at www.genomes-to-life.org.

# 14

## Improving Microarray Analysis with Hyperspectral Imaging, Experimental Design, and Multivariate Data Analysis

**David M. Haaland**[1] (dmhaala@sandia.gov), Jerilyn A. Timlin[1], Michael B. Sinclair[1], Mark H. Van Benthem[1], Michael R. Keenan[1], Edward V. Thomas[1], M. Juanita Martinez[2], Margaret Werner-Washburne[2], Brian Palenik[3], and Ian Paulsen[4]

[1]Sandia National Laboratories, Albuquerque, NM; [2]University of New Mexico, Albuquerque, NM; [3]Scripps Institution of Oceanography, La Jolla, CA; and [4]The Institute for Genomic Research, Rockville, MD

At Sandia National Laboratories, we are combining hyperspectral microarray scanning, efficient experimental designs, and a variety of new multivariate analysis approaches to improve the quality of data and the information content obtained from microarray experiments. Our approach is designed to impact the Sandia-led GTL team's investigation of *Synechococcus* for carbon sequestration. Current commercial microarray scanners use univariate methods to quantify a small number of dyes on printed microarray slides. We have developed a new hyperspectral microarray scanning system that offers higher throughput for each microarray slide by allowing the quantitation of a large number of dyes on each slide. The new scanner has demonstrated improved accuracy, precision, and reliability in quantifying dyes on microarrays and yields a higher dynamic range than possible with current commercial scanners. We will present the design of the new scanner, which collects the entire fluorescence spectrum from each pixel of the scanned microarray, and the use of multivariate curve resolution (MCR) algorithms to obtain pure emission spectra and corresponding concentration maps from the hyperspectral image data. The new scanner has allowed us to detect contaminating autofluorescence that emits at the same wavelengths as the reporter fluorophores on microarray slides. With the new scanner, we are able to generate relative concentration maps of the background, impurity, and fluorescent labels at each pixel of the image. Since the MCR generated concentration maps of the fluorescent labels are unaffected by the presence of background and impurity emissions, the accuracy and useful dynamic range of the gene expression data are both greatly improved. We will also demonstrate that the new scanner helps us understand a variety of artifacts that have been observed with microarrays scanned using two-color scanners. Artifacts include high background intensities, "black holes," dye separation, the presence of unincorporated dye, and contaminants that have led to the practice of intensity-dependent normalizations.

We will describe statistically designed microarray experimental approaches that we have used to identify and eliminate experimental error sources in the microarray technology. These statistically designed experiments have led to dramatic improvements in the quality and reproducibility of yeast microarray experiments. The lessons learned from yeast arrays will be applied directly to our GTL investigations of *Synechococcus* microarrays. In addition, new approaches with multivariate algorithms that incorporate error covariance of the arrays into the multivariate analysis of microarrays will be presented along with methods to evaluate the relative performance of various gene selection, classification, and multivariate fitting algorithms.

Evaluation of hybridization experiments with initial 250 gene *Synechococcus* microarrays will be presented along with graphical methods designed to facilitate understanding of the quality and repeatability of the *Synechococcus* microarray data. Work has recently begun on the whole genome *Synechococcus* microarray experi-

ments. cDNA arrays of 2496 genes from *Synechococcus* have been printed on glass slides. We will present the unique design of the whole genome arrays which includes six replicates for each gene each printed with a different pin to capture the true within array repeatability of gene expression. The arrays also include multiple positive controls, negative controls, blanks, and solvent spots in each block of the whole genome microarray. In addition, Arabidopsis promoter 70mers were printed on the four corners of each block to assist in positioning and reading the array. If available at the time of the workshop, gene expression results from the whole genome *Synechococcus* microarrays will be presented.

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 US Dep't of Energy's Genomics:GTL program (www.doegenomestolife.org) under project, "Carbon Sequestration in *Synechococcus Sp*.: From Molecular Machines to Hierarchical Modeling," (http://www.genomes2life.org/). This project also supported in part by a grant from the W. M. Keck Foundation.

# 15

## Multi-Resolution Functional Characterization of *Synechococcus* WH8102

**Nagiza F. Samatova***[1] (samatovan@ornl.gov), Andrea Belgrano[9], Praveen Chandramohan[1], Pan Chongle[1], Paul S. Crozier[2], Al Geist[1], Damian Gessler[9], Andrey Gorin[1], Jean-Loup Faulon[2], Hashim M. Al-Hashimi[7], Eric Jakobbson[4], Elebeoba May[2], Anthony Martino[2], Shawn Means[2], Rajesh Munavalli[1], George Ostrouchov[1], Brian Palenik[5], Byung-Hoon Park[1], Susan Rempe[2], Mark D. Rintoul[2], Diana Roe[2], Peter Steadman[9], Charlie E. M. Strauss[3], Jerilyn Timlin[2], Gong-Xin Yu[1], Maggie Werner-Washburne[10], Dong Xu[8], Ying Xu[6], and **Grant Heffelfinger**[2] (gsheffe@sandia.gov)

*Presenting author

[1]Oak Ridge National Laboratory, Oak Ridge, TN; [2]Sandia National Laboratories, Albuquerque, NM and Livermore, CA; [3]Los Alamos National Laboratory, Los Alamos, NM; [4]University of Illinois, Urbana Champaign, IL; [5]Scripps Institution of Oceanography, La Jolla, CA; [6]University of Georgia, Atlanta, GA; [7]University of Michigan, Ann Arbor, MI; [8]University of Missouri, Columbia, MO; [9]National Center for Genome Resources, Santa Fe, NM; and [10]University of New Mexico, Albuquerque, NM

Although sequencing of multi-megabase regions of DNA has become quite routine, it remains a big challenge to characterize all the segments of DNA sequence with various biological roles such as encoding proteins and RNA or controlling when and where those molecules are expressed. The primary difficulty is that function exists at many hierarchical levels of description, it has temporal and spatial connections that are difficult to manage, and functional descriptions do not correspond to well-defined physical models like biological structures defined by Cartesian coordinates for atoms. Results from the completed prokaryotic genome sequences show that almost half of the predicted coding regions identified are of unknown biological function. Specifically, the completed sequencing of *Synechococcus* WH8102 by JGI and multi-institutional annotation effort[1] has resulted in 1196 (out of 2522) ORFs that are conserved hypothetical or hypothetical.

The goal of this work is to develop a suite of computational tools for systematic multi-resolution functional characterization of microbial genomes and utilize them

to elucidate the biochemical mechanisms of the carbon sequestration of *Synechococcus sp.* as part of the SNL/ORNL GTL Center[2]. Functional characterization is conducted on many levels and with different questions in mind – ranging from the reconstruction of genome-wide protein-protein interaction networks to detailed studies of the geometry/affinity in a particular complex. Yet as the questions asked at the different levels are often intricately related and interconnected, we are approaching the problem from several directions. Here we outline some of them and provide pointers to more information.

While analysis of a single genome provides tremendous biological insights on any given organism, comparative analysis of multiple genomes can provide substantially more information on the physiology and evolution of microbial species. Comparative studies expand our ability to better assign putative function to predicted coding sequences and our ability to discover novel genes and biochemical pathways. The KeyGeneMiner[3] aims to identify "key" genes that are responsible for a given biochemical process of interest. When applied to the oxygenic photosynthetic process, it has discovered 126 genome features. Many of them have been reported in literature as either photosynthesis-related or photosynthesis-specific (occurring only in photosynthetic genomes). Likewise, the construction of comprehensive phylogenetic profiles for all transport proteins in the bacterial genomes[4] allows us to pick up regulators of transport and help annotate some genes for which there is still no, or weak, annotation. The approach is not limited to transport proteins; it can be done with any set of probe sequences representing an interesting functional grouping.

After genes are assigned to putative biochemical processes or putative functional links are established between genes, there still remains a significant challenge to understand how these biomolecules interact to form pathways for metabolic conversion from one substance to another and how genes form networks to regulate the timing and location events within the cell. In spite of some promising work using Boolean and Bayesian networks, all these approaches are challenged with too many parameters (compared to the number of data points) to adequately constrain the problem thus resulting in too many plausible competing gene models. The complexity of this problem is begging for novel methods that could integrate information from appropriate databases (e.g., gene expression, protein-protein interaction data, operon and regulon structure, transcription factors) in order to constrain the set of plausible solutions as well as to use this information for designing targeted experiments for study of specific network modules. Our progress towards this goal includes the development of new and effective protocols for systematic characterization of regulatory pathways and a preliminary version of a computational pipeline for interpretation of multiple types of biological data for biological pathway inference[5]. We have also prototyped these methods on *Synechococcus* WH8102 to make several predictions, including a signaling/regulatory network for the phosphorus assimilation pathway. We have developed an environment[6] to aid biologists in the analysis of proposed networks via visual browsing of the annotation information and original data associated with different elements of these networks. Using these tools and large visualization corridor with 48 high-resolution screens we have been able to begin assigning biological processes to groups of genes which were aggregated together by similar expression, as measured by microarrays, and then placed into a Boolean network based on discrete (Boolean) expression levels.

Structural characterization of protein machines provides additional valuable insights about the mechanism and details of their function. In spite of a long history behind the development of computational methods for structural characterization of protein machines, many challenges still remain to be addressed. Specifically, we are focusing on the methods for understanding how biological molecules interact physi-

cally to transfer signals, including protein-protein interactions as well as protein-DNA and protein-RNA interactions[7]. At the initial stage we apply structure prediction methods to determine protein fold families with our ROSETTA[8] and PROSPECT[9] programs and use inferred structural similarities to create hypotheses about their interacting partners. The computational pipeline merges several bioinformatics and modeling tools including algorithms for protein domain division, secondary structure prediction, fragment library assembly, and structure comparison[7]. Application of this pipeline has resulted in genome-scale structural models for *Synechococcus* genes[10].

Protein interactions with other molecules play a central role in determining the functions of proteins in biological systems. Protein-protein, protein-DNA, protein-RNA, and enzymes-substrate interactions are a subset of these interactions that is of key importance in metabolic, signaling and regulatory pathways. Bacterial chemotaxis, osmoregulation, carbon fixation and nitrogen metabolism are just a few examples of the many complex processes dominated by such molecular recognition. Identification of interacting proteins is an important prerequisite step in understanding its physiological function. We develop several complimentary approaches to this problem including statistically significant protein profiles based[3], signature kernel SVMs based[11], and genomic context based[5]. Utilization of these approaches produced a genomes-scale protein interaction map for *Synechococcus* WH8102.

Knowing interacting partners is just the first step towards elucidating the order and control principles of molecular recognition. One of the most remarkable properties of protein interactions is high specificity. Even presumably specific binding sites may bind a range of ligands with different compositions and shapes. For example, the Ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco) can catalyze two separate reactions, carboxilation and oxygenation reactions depending on whether $CO_2$ or $O_2$ binds in its active site, respectively. There must be then *functionally important residues* that enable different proteins to recognize their unique interacting partners. Identification of these functionally important residues (e.g. docking interfaces, catalytic centers, substrate and cofactor binding sites and hinge-motion controlling loops) is essential for functional characterization of protein machines. We explore several complimentary approaches to identify protein docking interfaces based on quantification of correlated mutations[3], Boosting driven separation into "predictable" feature subspaces[3], and statistically significant separation of likely *n*-mers[6]. Likewise, our Surface Patch Ranking method[3] has been utilized to identify clusters of residues important for $CO_2/O_2$ specificity in Rubisco. Finally, our tools for full atom modeling of protein interactions (LAMMPS, PDOCK)[10] are being used to assess Rubisco catalytic specificity.

All these tools and experiments generate heterogeneous data (e.g., sequence, structure, biochemical, physiological and genetic data). Integrative analysis of these data will yield major insights into the carbon sequestration behavior of *Synechococcus sp*. For this to occur, low level molecular data and processes need to be connected to macro-ecological models. We are doing this in a Hierarchical Simulation Platform[12] that uses recently published models connecting metabolic rate to population growth and trophic level energy flux. Additionally, we have built a web portal for *Synechococcus* Encyclopedia[13] that enables an efficient access, management, curation, and computation with these data as well as comparative analysis with other microbial genomes.

**References**

1. http://genome.ornl.gov/microbial/syn_wh

2. Grant Heffelfinger (gsheffe@sandia.gov) and Al Geist (gst@ornl.gov); http://www.genomes2life.org

3. Nagiza Samatova (samatovan@ornl.gov)

4. Eric Jakobsson (jake@ncsa.uiuc.edu)

5. Ying Xu (xyn@ornl.gov), Brian Palenik (bpalenik@ucsd.edu), and Dave Haaland (dmhaala@sandia.gov); see "Microarray Analysis with Hyperspectral Imaging, Experimental Design, and Multivariate Data Analysis" and "Methods for Ellucidating Synechococcus Regulatory Pathways" posters

6. George S Davidson (gsdavid@sandia.gov)

7. Andrey Gorin (agor@ornl.gov); see also "Bioinformatics Methods for Mass Spect Analysis" poster

8. Charlie Strauss (cems@lanl.gov)

9. Ying Xu (xyn@ornl.gov) and Dong Xu (xudong@missouri.edu); http://compbio.ornl.gov/structure/prospect/

10. Ying Xu (xyn@ornl.gov); http://compbio.ornl.gov/PROSPECT/syn/

11. Daniel M Rintoul (mdrinto@sandia.gov) and Antony Martino (martino@sandia.gov); see also "Modeling Cellular Response" poster

12. Damian Gessler (ddg@ncgr.org), Andrea Belgrano (ab@ncgr.org), and Peter Steadman (ps@ncgr.org).

13. Al Geist (gst@ornl.gov) and Nagiza Samatova (samatovan@ornl.gov); http://www.genomes-to-life.org/syn_wh; see also "The *Synechococcus* Encyclopedia" poster.

# 16

## Computational Inference of Regulatory Networks in *Synechococcus sp* WH8102

Zhengchang Su[1], Phuongan Dam[1], Hanchuan Peng[1], **Ying Xu**[1] (xyn@bmb.uga.edu), Xin Chen[2], Tao Jiang[2], Dong Xu[3], Xuefeng Wan[3], and Brian Palenik[4]

[1]University of Georgia, Athens, GA and Oak Ridge National Laboratory, Oak Ridge, TN; [2]University of California, Riverside, CA; [3]University of Missouri, Columbia, MO; and [4]University of California, San Diego, CA

In living systems, control of biological function occurs at the cellular and molecular levels. These controls are implemented by the regulation of activities and concentrations of species taking part in biochemical reactions. The complex machinery for transmitting and implementing the regulatory signals is made of a network of interacting proteins, called *regulatory networks*. Characterization of these regulatory networks or pathways is essential to our understanding of biological functions at both molecular and cellular levels.

We have been developing a prototype system for computational inference of regulatory and signaling pathways for the genome of *Synechococcus sp.* WH8102. Currently, the prototype system consists of the following components: (a) prediction of gene

functions, (b) prediction of terminators of operon structures, (c) genome-scale prediction of operon structures, (d) genome-scale prediction of regulatory binding sites, (e) mapping of orthologous genes and biological pathways across related microbial genomes, (f) prediction of protein-protein interactions, (g) mapping biological pathways across related genomes, and (h) inference of pathway models through fusing the information collected in steps (a) through (g).

a. **Computational prediction of gene functions.** We have previously developed a computational pipeline for inference of protein structures and functions at genome scale (Shah, et al. 2003 and Xu, et al. 2003). The pipeline consists of both sequence-based homology detection programs like psi-BLAST, and structure-based homology detection program PROSPECT (Xu, et al. 2000). We found that using structure-based approach in addition to psi-BLAST, we can detect additional 10-20% of remote homologs for genes in a microbial genome. This pipeline can be accessed at http://compbio.ornl.gov/PROSPECT/PROSPECT-Pipeline/cgi-bin/proteinpipeline_form.cgi. We have applied this pipeline to all the orfs of *Synechococcus sp* WH8102, and assigned close to 80% of its genes to some level of functions. All the results can be found at http://compbio.ornl.gov/PROSPECT/syn/.

b. **Prediction of terminators of operons.** We are in the early phase of developing a computational capability for prediction of terminators in WH8102. Our initial focus has been on *rho*-independent terminators (RIT). We are carrying out a comparative genome analysis to compare the RITs of the otthologous genes in different genomes to identify possible conserved patterns. We are also developing and implementing a novel algorithm based on MST clustering approaches to use common features of RITs to predict new RITs. We will apply more sophisticated energy functions than the one used in TransTerm and RNAMotif.

c. **Prediction of operons.** We have been working on a comparative genomics approach for predicting operons in *Synechococcus sp*. WH8102 that combines many known characteristics of an operon structure concerning the functions, intergenic distances and transcriptional directions of genes, promoters, terminators, etc. in a unified likelihood framework (Chen, et al. 2003). The data and results are available to the public at http://www.cs.ucr.edu/~xinchen/operons.htm. We have used the predicted operons, as one piece of information, in our inference of regulatory pathways in *Synechococcus sp* WH8102.

d. **Prediction of regulatory binding sites.** We have previously developed a computer program CUBIC for identification of consensus sequence motifs as possible regulatory binding sites (Olman et al. 2003). CUBIC solves the binding site identification problem as a problem of identifying data clusters from a noisy background. We have applied CUBIC for binding site predictions at genome scale, through identifying orthologous genes of WH8102 in other related genomes and application of CUBIC to the upstream regions of each set of orthologous genes. We expect that the genome-scale binding site prediction results will be publicly available within weeks.

e. **Mapping of orthologous genes across related genomes.** The identification of orthologous genes is a fundamental problem in comparative genomics and evolution, and is very challenging especially on a genome-scale. We have been working on anew approach for assigning orthologs between different (but related) genomes based on homology search and genome rearrangement. The preliminary experimental results on simulated and real data demonstrate that

the approach is very promising (it is competitive to the existing methods), although more needs to be done (Chen, et al. 2004).

**f. Prediction of protein-protein interactions.** We have implemented a computer software for predicting protein-protein interactions, employing a number of popular prediction strategies, including mapping against protein-protein interaction maps derived from experiments (like two hybrid), application of phylogenetic profile analysis (Pellegrini et al. 1999) and gene fusion method (Marcotte et al. 1999). We have made a genome-scale prediction of protein-protein interactions for *Synechococccus sp* genes.

**g. Mapping of biological pathways across related genomes.** We have recently developed a computational method for mapping of biological pathways across related microbial genomes. The core component of the algorithm/program is orthologous gene mapping under the constraints of (a) operon structures, (b) regulon structures (defined in terms of operons with common regulatory binding sites), and (c) co-expressions of genes. We have implemented this algorithm as a computer program P-MAP, and apply this program to assign all known pathways in *E coli*. (partial or complete) to *Synechococcus sp*. WH8102. The mapping results will soon be posted at our *Synechococcus* Knowledge Dabase at http://csbl.bmb.uga.edu/~peng/home.html.

**h. Pathway inference through information.** We have developed a computational protocol for inference of regulatory and signaling pathways through fusing the information collected in the steps (a) through (g) and a simple merging-voting scheme to put the predicted complexes, protein-DNA interactions and protein-protein interactions. We have applied this capability to the inference to the prediction of a number of regulatory pathways, including phosphorus assimilation pathway [ref], nitrogen and carbon assimilation pathways [unpublished results] of *Synechococcus sp* WH8102. We are currently exploring a number of formalisms for piecing together predicted gene associations into pathway models, including Biochemical Systems Theory (BST) (Savageau 1976).

Experimental validations of predictions are being carried out using microarray analyses (see Haaland et al poster) of wild type WH8102 and knockout mutants under selected growth conditions.

**ACKNOWLEDGEMENTS.** This work is funded in part by the US Department of Energy's Genomics:GTL (www.doegenomestolife.org) under project "Carbon Sequestration in *Synechococcus sp*: From Molecular Machines to Hierarchcal Modeling" (www.genomes-to-life.org).

## References

1. X. Chen, Z. Su, P. Dam, B. Palenik, Y. Xu, and T. Jiang. Operon prediction by comparative genomics: an application to the Synechochoccus sp. WH8102 genome. 2003, submitted to *Nuc. Acids Res*. (in revision).

2. X. Chen, J. Zheng, Z. Fu, P. Nan, Y. Zhong, S. Lonardi, and T. Jiang. Assignment of orthologous genes via genome rearrangement. 2004, submitted to ISMB'2004.

3. E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice et.al, Detecting protein function and protein-protein interactions from genome sequences, *Science*, **285**:751-753, 1999.

4. V. Olman, D. Xu and Ying Xu, "Identification of Regulatory Binding-sites using Minimum Spanning Trees", Proceedings of the 7th Pacific Symposium on Biocomputing (PSB), pp 327-338, 2003.

5.

M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg et.al, Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, *Proc Natl Acad Sci USA*, **96**:4285-4288, 1999.

6. M. Shah, S. Passovets, D. Kim, K. Ellrott, L. Wang, I. Vokler, P. Locascio, D. Xu, Ying Xu, A Computational Pipeline for Protein Structure Prediction and Analysis at Genome Scale, *Proceedings of IEEE Conference on Bioinformatics and Biotechnology*, 3-10, IEEE/CS Press, 2003 (An expanded journal version is published in *Bioinformatics*, **19**(15):1985-1996, 2003).

7. M. A. Savageau. "Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology," Addison-Wesley, Reading, Mass (1976).

8. Z. Su, A. Dam, X. Chen, V Olman, T. Jiang, B. Palenik, and Ying Xu, Computational Inference of Regulatory Pathways in Microbes: an application to the construction of phosphorus assimilation pathways in Synechococcus WH8102, Proceedings of 14th *International Conference on Genome Informatics* pp:3-13,Universal Academy Publishing, 2003.

9. D. Xu, P. Dam, D. Kim, M. Shah, E. Uberbacher, and Ying Xu, Characterization of Protein Structrue and Function at Genome-scale using a Computational Prediction Pipeline, accepted to appear in *Genetic Engineering: Principles and Methods*, Vol **32**, Jane Setlow (Ed.), Plenum Press, 2003.

10. Y. Xu and D. Xu. Protein threading using PROSPECT: Design and evaluation. *Proteins: Structure, Function, and Genetics*. **40**:343-354. 2000.

## University of Massachusetts, Amherst

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

# 17

## Analysis of Predominant Genome Sequences and Gene Expression During *In Situ* Uranium Bioremediation and Harvesting Electricity from Waste Organic Matter

Stacy Ciufo[1]*, Dawn Holmes[1], Zhenya Shelbolina[1], Barbara Methé[2], Kelly Nevin[1], and **Derek Lovley**[1] (dlovley@microbio.umass.edu)

*Presenting author

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

Field studies have demonstrated that stimulating dissimilatory metal reduction in uranium-contaminated subsurface environments is an effective, simple, and inexpensive method for removing uranium from contaminated groundwater. Molecular analyses, which avoid any culture bias, have demonstrated that *Geobacter* species are the predominant microorganisms in a variety of subsurface environments in which dissimilatory metal reduction is an important process. For example, during *in situ* bioremediation of a uranium-contaminated site in Rifle, Colorado *Geobacter* species accounted for as much as 80% of the microbial community in the groundwater during uranium bioremediation. In a similar manner, it has been demonstrated that *Geobacter* species are the predominant microorganisms on electrodes used to harvest electricity from waste organic matter. In order to determine whether models of *Geobacter* physiology derived from pure culture studies are applicable to as-yet-uncultured *Geobacter*s living in uranium-contaminated subsurface environments or on the surface of electrodes, it is necessary to determine the relative similarities of the genome sequence and gene expression patterns of as-yet-uncultured *Geobacter*s and the pure cultures.

One strategy to evaluate the genetic potential of the *Geobacter*s that predominate in the environments of interest is to sequence genomic DNA directly extracted from the environment. Genomic DNA was extracted from the sediments of a uranium-contaminated aquifer, located in Rifle, Colorado, in which the activity of *Geobacter*s had been stimulated with the addition of acetate to promote precipitation of uranium. The DNA was cloned in bacterial artificial chromosomes (BACs) with an average insert size of ca. 40 kbp. Large scale sequencing of the BAC inserts resulted in the recovery of 4.2 mbp of environmental genomic DNA sequence. The sequence data was assembled with BACPACK, an algorithm we specifically developed for this purpose. A contig of over 580 kbp of sequence was assembled as were

several other contigs of 25-475 kbp. Analysis for highly conserved *Geobacter* genes indicated that these contigs were from a *Geobacter* species. The uncultured *Geobacter* had a 16S rRNA gene sequence identical to a sequence that predominated in the groundwater during uranium bioremediation. The environmental genome sequence was similar to that of pure *Geobacter* species in that it had a lower percentage of putative proteins predicted to be localized in the cytoplasm and more proteins targeted to the inner membrane, periplasm, and outer membrane than has been found in non-*Geobacteraceae*. The uncultured *Geobacter* species had a high percentage of *Geobacter* signature genes and gene arrangements similar to those found in cultured species. However, there were also genes of unknown function in the uncultured *Geobacter* that have not been identified in the genomes of any pure cultures. These results suggest that although the uncultured *Geobacter* species involved in uranium bioremediation are clearly not identical to the pure cultures that are being intensively studied, there are many genomic similarities and thus models of pure cultures may have applicability to *Geobacter*-dominated subsurface environments.

Another strategy to determine the genome sequences of the *Geobacter*s predominating in environments of interest is to adapt culture conditions to permit culturing of these organisms. A medium in which the clay-fraction of subsurface sediments was the source of Fe(III) oxide was developed. With this medium a *Geobacter* species with a 16S rRNA gene sequence identical to one of the sequences that predominated during uranium bioremediation was isolated. Cultivation of this organism required the addition of groundwater from the site to the medium. Sufficient quantities of this organism have now been cultured so that its genome can be sequenced. Of further interest is the finding that one of the large BAC contigs from the uranium bioremediation site has the same 16S rRNA gene sequence. Thus, it will be possible to compare results from direct sequencing of environmental genomic DNA with the strategy of isolation in culture followed by genome sequencing.

One test of the environmental applicability of the current physiological models of *Geobacter* metabolism is to determine whether *Geobacters* in environments of interest have patterns of gene expression that are similar to those in pure cultures. Therefore, methods for effectively extracting mRNA from aquifer sediments and the surface of energy-harvesting electrodes were developed. Initial studies on the metabolic state of *Geobacter* species in aquifer sediments demonstrated that the natural populations of *Geobacters* were highly expressing genes for nitrogen fixation, suggesting that they were limited for fixed nitrogen. Addition of 100 $\mu$M ammonium to the sediment repressed expression of the nitrogen fixation genes. These results demonstrated that it is possible to evaluate the *in situ* metabolic state of *Geobacters* in subsurface environments. The next step will be to evaluate the expression of a larger suite of genes involved in nutrient uptake and stress response with microarrays.

Evaluation of gene expression in *Geobacter sulfurreducens* growing on the surface of energy-harvesting electrodes suggested that environmental analysis with whole-genome DNA microarrays is feasible. A microarray was used to compare mRNA levels of *G. sulfurreducens* growing on electrodes with mRNA levels of planktonic cells. Up-regulation of several genes was significant on the electrode. For example, mRNA levels for several outer-membrane cytochromes were 40-80 fold higher in cells growing on the electrodes. This suggests that these cytochromes play an important role in electron transfer to the electrode surfaces. There was also an upregulation of genes annotated as encoding for heavy-metal efflux proteins. This may reflect the presence of heavy metal contaminants in the electrode material. Many genes of unknown function were also down-regulated. The most prominently down-regulated genes were related to oxygen respiration and/or oxygen toxicity. These included a cytochrome oxidase as well as thioredoxin peroxidase and

superoxide dismutase. This is indicative of an important change in the respiratory pathway. These results demonstrate that electron transfer to electrodes is associated with significant shifts in gene expression and provide the first insights into the mechanisms for this novel form of respiration.

In summary, these initial results demonstrate that it will be possible not only to determine the genetic potential of the *Geobacter* species actually involved in subsurface bioremediation or in harvesting electricity from waste organic matter, but also to broadly assess their metabolic state. This will significantly improve the development of *in silico* models for predicting the metabolic responses of *Geobacter* species under different environmental conditions and provide information on how to most effectively optimize these applications of *Geobacter*.

# 18

## Functional Analysis of Genes Involved in Electron Transport to Metals in *Geobacter sulfurreducens*

Maddalena Coppi[1]*, Eman Afkar[1], Tunde Mester[1], Daniel Bond[1], Laurie DiDonato[1], Byoung-Chan Kim[1], Richard Glaven[1], Ching Leang[1], Winston Lin[1], Jessica Butler[1], Teena Mehta[1], Susan Childers[1], Barbara Methé[2], Kelly Nevin[1], and **Derek Lovley**[1] (dlovley@microbio.umass.edu)

*Presenting author

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

As noted in a companion abstract, ecological studies have demonstrated that *Geobacter* species are the predominant microorganisms in a variety of subsurface environments in which dissimilatory metal reduction is an important process, including during *in situ* uranium bioremediation. Therefore, in order to effectively model *in situ* bioremediation of uranium and develop strategies for improving this process it is necessary to understand the factors controlling the growth and activity of *Geobacter* species. Most important in this regard is information on electron transfer not only to U(VI), but also to Fe(III), because most of the energy supporting the growth of *Geobacter* species in uranium-contaminated subsurface environments is derived from electron transfer to Fe(III).

Functional analysis of electron transfer to metals in *Geobacter sulfurreducens* has initially focused on the *c*-type cytochromes which are abundant in the genome, as well as outer-membrane proteins of previously unknown function. For example, analysis of the outer-membrane proteins of *G. sulfurreducens* with MALDI-TOF mass spectrometry revealed that the most abundant protein, designated OmpA, had a predicted amino acid sequence without any significant homology with previously described genes. OmpA is predicted to have a hydrophobic leader sequence, consistent with export to the outer membrane, and a β-barrel structure. No heme *c* or metal binding motifs were detected. When the gene was deleted with the single gene replacement method, the *ompA*-deficient mutant grew the same as wild type with fumarate as the electron acceptor, but it could not grow with Fe(III) or Mn(IV) oxides as the electron acceptor. Although the total heme *c* content in the mutant and the wild type were comparable, the mutant had only ca. 50% of the heme *c* content in the outer membrane as the wild type. There was a corresponding substantial increase in the total heme *c* content in the cytoplasmic membrane and

soluble fraction of the *ompA* mutant. These results suggest that OmpA plays an important role in localizing *c*-type cytochromes in the outer membrane of *G. sulfurreducens* via a novel mechanism not previously described in any microorganism.

A mutation in a novel secretory system in *G. sulfurreducens* specifically eliminated its ability to reduce Fe(III) oxides, but not soluble electron acceptors, including chelated Fe(III). Comparison of the proteins in the periplasm of this mutant with wild type cells indicated that several proteins were accumulating in the periplasm of the mutant. Analysis of peptide fragments of one of these proteins revealed a gene, designated *ompB*, which encodes for a 1303 amino acid protein, with 23 transmembrane amino acids and 1275 amino acids predicted to be exposed outside the cell. There are four putative metal-binding sites. This gene is found in the four *Geobacteraceae* genome sequences that are available, but not in any other organisms. The *ompB* mutant did not grow on Fe(III) oxide, but grew on soluble electron acceptors. These results suggest that OmpB plays an important role in cell-Fe(III) oxide contact or in sequestering Fe(III) from Fe(III) oxides, prior to Fe(III) reduction. This is a novel concept for dissimilatory Fe(III) oxide reduction.

Our previous studies have suggested that *c*-type cytochromes are important in electron transfer to Fe(III) in *Geobacter sulfurreducens*. However, elucidating which cytochromes are involved in Fe(III) reduction is not trivial because the genome of *G. sulfurreducens* contains genes for over 100 *c*-type cytochromes, at least 25 of which are predicted to be localized in the outer membrane where Fe(III) reduction is likely to take place. Therefore, our initial strategy has been to focus on cytochromes predicted to be localized in the outer membrane, as well as cytochromes that are specifically expressed during growth on Fe(III). Functional analysis of nearly all of the outer-membrane *c*-type cytochromes has led to the surprising result that, in many instances, the deletion of just one of the cytochrome genes severely inhibits Fe(III) reduction. This suggests that many of the multiple outer-membrane cytochromes do not serve duplicative functions, but act in concert to bring about electron transfer to Fe(III).

There are also many periplasmic cytochrome genes that are highly similar. Mutants were generated in order to evaluate their role in electron transfer to metals. It was found that PpcA, PpcB, and PpcC are required for soluble Fe(III) reduction but that mutants that could no longer produce PpcD or PpcE grew better on Fe(III) than the wild type, as did a double mutant lacking PpcB and PpcC. These results demonstrate that despite their apparent similarities in size and heme content, these periplasmic cytochromes have some different functions in electron transfer in *G. sulfurreducens* and that it is possible to make mutations that will enhance electron transfer to metals.

It has been proposed that, based on analogy to our previous findings in *Desulfovibrio vulgaris*, *c*-type cytochromes are also important electron carriers for U(VI) reduction. Analysis of over 15 *c*-type cytochrome mutants suggested that the small periplasmic *c*-type cytochromes in *G. sulfurreducens*, which are most closely related to the $c_3$ cytochrome responsible for U(VI) reduction in *D. vulgaris*, were not responsible for U(VI) reduction. However, knockout mutations in several outer-membrane cytochromes inhibited U(VI) reduction. These results suggest that the mechanisms for U(VI) reduction in *G. sulfurreducens* are significantly different than for *D. vulgaris* and indicate that even though U(VI) is soluble, and could potentially be reduced in the periplasm, reduction by *G. sulfurreducens* is more likely to take place primarily at the outer membrane surface.

Sequencing of the *G. sulfurreducens* genome revealed the presence of genes predicted to be involved in oxygen respiration, which was surprising because no *Geobacter* species had ever been found to grow on oxygen. However, growth conditions under which *G. sulfurreducens* can grow at oxygen concentrations that are 50% or less of atmospheric levels have now been identified. Knockout mutation studies demonstrated that growth on oxygen is dependent upon a cytochrome oxidase. The ability of *Geobacter* species to grow at low oxygen levels helps explain how they survive in aerobic subsurface environments and then rapidly respond to the development of anaerobic conditions during metals bioremediation.

Functional analysis of proteins important in central metabolism, such as a novel eukaryotic-like citrate synthase and a bifunctional succinate dehydrogenase/fumarate reductase, has also been completed. These studies are rapidly improving the understanding the physiology of *G. sulfurreducens*. This information will permit more informed decisions on strategies to optimize bioremediation and energy harvesting applications of *Geobacter* species.

# 19

## Adapting Regulatory Strategies for Life in the Subsurface: Regulatory Systems in *Geobacter sulfurreducens*

Gemma Reguera[1]*, Cinthia Nunez[1], Richard Glaven[1], Regina O'Neil[1], Maddalena Coppi[1], Laurie DiDonato[1], Abraham Esteve-Nunez[1], Barbara Methé[2], Kelly Nevin[1], and **Derek Lovley**[1] (dlovley@microbio.umass.edu)

*Presenting author

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

As outlined in accompanying abstracts, *Geobacter sulfurreducens* serves as a pure culture model for the *Geobacter* species that are responsible for *in situ* uranium bioremediation in contaminated subsurface environments and that harvest electricity from waste organic matter. In order to predictively model the activity of *Geobacter*s involved in bioremediation and energy harvesting it is necessary to understand how electron transport to metals as well as central metabolism are regulated under different environmental conditions.

Of particular relevance for bioremediation and energy harvesting applications of *Geobacter* species is understanding regulation of gene expression under the sub-optimal growth conditions typically encountered in subsurface environments. For example, the genome of *G. sulfurreducens* contains a homolog of the *E. coli* stationary-phase sigma factor, RpoS, which is of interest because growth in the subsurface is likely to be analogous to the stationary phase of cultures. Survival in stationary phase, aerotolerance, growth on oxygen, and reduction of insoluble Fe(III) were diminished in an *rpoS* mutant, but there was no apparent impact on response to high temperature or alkaline pH stress, as seen in *E. coli*. In order to further elucidate the rpoS regulon, gene expression in the *rpoS* mutant and the wild type were compared with whole genome DNA microarray and proteomics approaches. These studies demonstrated that RpoS controls genes involved in Fe(III) reduction, oxygen tolerance, and oxygen respiration. This study represents the first characterization of RpoS in a member of the δ subclass of the *Proteobacteria*

and suggests that RpoS plays an important role in regulating metabolism of *Geobacter* species under the stressful conditions found in subsurface environments.

RpoS negatively regulates another sigma factor, RpoE, which modulates a distinct regulon also involved in oxygen tolerance and repair of oxidative stress damage. RpoE also was found to have an important role in controlling attachment to Fe(III) oxide and electrode surfaces, two key processes for the environmental success of *Geobacter* species. Genome-wide transcriptional profiles of *G. sulfurreducens* biofilms grown on Fe(III) oxide surfaces versus their planktonic counterparts, as well as transcriptional profiling of an *rpoE* mutant, suggested that RpoE regulates the transition from planktonic to biofilm conditions as well as maintenance of the biofilm mode of growth and electron transfer to Fe(III). These results demonstrate that RpoE and RpoS act coordinately to finely tune the adaptive responses that enable *Geobacters* to survive and outcompete many other organisms in subsurface environments.

RelA is another regulatory protein that could be important in influencing growth in the subsurface as a mutant in the putative *relA* gene in *G. sulfurreducens* grew faster than wild type under nutrient limitation. Microarray analyses of the *relA* mutant demonstrated that, as in *E. coli*, ribosomal proteins and chaperones are negatively regulated by RelA, while stress response genes are positively controlled, further suggesting that RelA may play a critical role in slow growth and stress response. In addition, RelA also appeared to positively regulate proteins required for the reduction of insoluble Fe(III) reduction, thus illustrating that in *G. sulfurreducens* RelA has unique targets that link the regulation of growth rate to metal reduction.

Analysis of the *G. sulfurreducens* genome revealed that this organism is highly attuned to its environment with 5.2% of the open reading frames in the genome dedicated to two-component proteins and 1.9% dedicated to chemotaxis. A combination of genomics and proteomics approaches identified putative histidine kinases and response regulators and results of microarray analyses of wild type and selected mutants enabled preliminary characterization and pairing of 16 two-component signal transduction proteins that previously were of unknown function and classified as "orphans". Histidine kinase knockouts in *G. sulfurreducens* over-expressing the cognate response regulators produced information on the environmental signals triggering regulatory cascades and provided further support for the role of two component systems in integrating responses to environmental stimuli with electron transfer.

*Geobacter* species generally live in environments high in dissolved Fe(II) and have unusually high requirements for iron due to their high cytochrome content. In *G. sulfurreducens*, concentrations of dissolved Fe(II) as high as 100 $\mu$M were found to be required for optimal growth and acetate uptake and the cellular iron content greatly exceeded that of *E. coli*, suggesting that mechanisms to regulate iron uptake and iron overload in *Geobacters* may be different than in other, previously studied organisms. A homolog of the *E. coli* Fe(II)-dependent ferric uptake regulator, Fur, was identified in the *G. sulfurreducens* genome. As in *E. coli*, expression of *fur* in *G. sulfurreducens* was repressed in the presence of Fe(II) and the phenotype of a *fur*-knockout mutant suggested that Fur has a key role in responding to changes in Fe(II) concentration in the environment. Only a small fraction of Fur-regulated genes identified by microarray analysis were preceded by a recognizable Fur box. Surprises in the genes under Fur control included proteins required for Fe(III) oxide reduction.

These studies, as well as other ongoing studies on novel regulatory strategies in *G. sulfurreducens*, suggest that models of regulation that have been developed in previ-

ously studied microorganisms can help in identifying some of the regulatory components in *G. sulfurreducens* but, in many instances, regulation patterns and mechanisms in *G. sulfurreducens* have been modified in order to adapt to life in the subsurface. Results from these regulation studies will be incorporated into the expanding *in silico* model of *G. sulfurreducens* in order to better predict the likely response of *Geobacter* species during attempts to optimize bioremediation and energy harvesting strategies.

## See Also

- *In Silico* Elucidation of Transcription Regulons and Prediction of Transcription Factor Binding Sites in *Geobacter* Species Using Comparative Genomics and Microarray Clustering, Krushkal et al, on page 73.

## *Shewanella* Federation

# 20

## Global and Physiological Responses to Substrate Shifts in Continuous and Controlled Batch Cultures of *Shewanella oneidensis* MR-1

**Jim Fredrickson** (jim.fredrickson@pnl.gov), Alex Beliaev, Bill Cannon, Yuri Gorby, Mary Lipton, Peter Liu, Margie Romine, Richard Smith, and Harold Trease

Pacific Northwest National Laboratory, Richland, WA

Collaborating *Shewanella* Federation Team Leaders: Carol Giometti (Argonne NL); Eugene Kolker (BIATECH); Ken Nealson (USC); Monica Riley (MBL); Daad Saffarini (UW-M); Jim Tiedje (MSU), and Jizhong Zhou (Oak Ridge NL)

*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 Responses**. To define the networks of genes responding to metal electron acceptors, mRNA expression patterns of cells reducing fumarate were compared to those reducing nitrate, thiosulfate, DMSO, TMAO, and several forms of Fe(III) and Mn(III) using whole-genome arrays of *S. oneidensis*. Analysis of variance performed on the complete dataset identified over 1600 genes displaying significant expression changes across different metal-reducing conditions. Two principal components accounted for 78% of the variability within the multiple-electron-acceptor dataset and were represented by genes displaying specific response to metals. Hierarchical clustering revealed a high degree of similarity in mRNA relative abundance levels was displayed for all the metal-reducing conditions; all clustering separately from the inorganic electron acceptors. Interestingly, no significant differences in expression profiles were observed between solid and soluble metal acceptors. Only a few genes specific for any particular metal were identified. In contrast, K-means clustering identified a group of over 150 genes displaying highly specific up-regulation under all metal-reducing conditions. Among those, we identified putative transporters, outer membrane components, as well as two electron transfer proteins (flavodoxin and a *c*-type cytochrome). Further work will be aimed at differentiating cells responses to divalent metal cations (i.e., reduction products) and the oxidized form of the metals; and functional characterization of the differentially regulated genes.

**Response to $O_2$ Concentrations.** Autoaggregation occurs in *Shewanella oneidensis* MR-1 cultures growing at high $O_2$ concentrations in the presence of $Ca^{2+}$ ions. 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. In an effort to elucidate these mechanisms and identify processes associated with $O_2$-induced autoaggregation in *S. oneidensis*, a comparative analysis using DNA microarrays was performed on samples grown under different $O_2$ tensions in the presence and absence of $Ca^{2+}$. Although, when compared to $O_2$-limited conditions, both flocculated and unflocculated cells displayed some similarities in gene expression in response to elevated levels of $O_2$, including genes involved in cell envelope functions and EPS/LPS production, autoaggregation had a significant impact on gene expression in MR-1. Direct comparison of aggregated versus nonagreggated cells grown under 50% dissolved $O_2$ tension (DOT) revealed remarkable differences in mRNA patterns between these two states. The nonaggregated cells displayed significant increase of mRNA levels of genes involved in aerobic energy metabolism, amino acid and cofactor biosynthesis, as well as chemotaxis and motility. In contrast, genes putatively involved in anaerobic metabolism (fumarate and polysulfide reductases, and Ni/Fe hydrogenase), cell attachment (type IV pilins and curli), and transcription regulation (*rpoS*, *spo*IIAA) were upregulated under 50% DOT aggregated conditions. Notably, a gene cluster encoding outer membrane proteins and cytochromes (*mtrDEF*) also displayed up to 7-fold increase in mRNA levels in aggregated cells. 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.

**Carbon Metabolism**. In contrast to the wide array of electron acceptors reduced by *S. oneidensis*, this organism is relatively limited in regards to utilization of multicarbon substrates for anaerobic respiration. Earlier studies indicated that MR-1 can utilize formate as a sole source of carbon and energy under anaerobic condition. A hypothetical amalgam pathway for lactate metabolism that included the elements of serine-isocitrate cycle for formate utilization was previously proposed by K. Nealson and colleagues. The availability of whole-genome sequence allowed compilation of a possible pathway for formate assimilation in *S. oneidensis.* MR-1 is predicted to possess a number of putative enzymes including pyruvate formate-lyase (PFL) that may allow for the assimilation of both exogenous and lactate-derived formate through a modification of a serine-isocitrate pathway. Three independent lines of evidence of methylotroph-like metabolism of lactate in *S. oneidensis* MR-1 are provided: lactate is stoichiometrically converted to acetate in $O_2$-limited or lactate-excess anaerobic chemostat cultures and no formate is present in the supernatant; the amount of PFL protein increased 2.5-fold under $O_2$ limited growth compared to fully aerobic growth; and relative abundance of mRNA from genes encoding key enzymes of the proposed pathway for formate assimilation including isocitrate lyase, malate synthase, and serine hydroxymethyl-transferase increased under $O_2$-limitation compared to aerobic growth. Moreover, biomass yield from $O_2$-limited or anaerobic chemostat cultures of MR-1 grown under excess lactate indicated that formate was utilized as the sole source of carbon. These results suggest the presence of an unusual mechanism of carbon metabolism of lactate in *S. oneidensis* with formate as the key element of the intermediary carbon metabolism under $O_2$-depleted and/or lactate excess conditions.

**Protein Secretion.** In many bacteria, translocation of key respiratory enzymes is mediated by the twin arginine translocation (TAT) machinery. Analysis of the N-terminal sequences of MR-1 proteins for conserved TAT leader properties revealed 30 candidates. This list includes 11 genes predicted to be responsible for utilization of

formate and $H_2$ as electron donors or for catalyzing the terminal step in electron transfer to nitrate, nitrite, and sulfur-containing substrates. Various uncharacterized proteases and putative redox active proteins were among the remaining 19 candidates. In order to investigate the role of secreted proteins in MR-1 respiratory metabolism, two mutants were constructed: a *tatC* gene deletion mutant and a transposon mutant of the terminal branch of type II general secretion pathway (*gspD*). The TAT mutant was unable to reduce metals with formate or $H_2$ as an electron donor. Although cells could grow with lactate and fumarate, there was a longer lag phase relative to the wild type. As expected, the ability to grow on nitrate and DMSO was abolished. Reduction of technetium ($TcO_4^-$) with lactate was severely impaired in the TAT mutant, likely due to mislocalization of one or more hydrogenases. Hydrogenase have reported to be involved in Tc reduction in other bacteria. The TAT system is a key cellular machine in MR-1 that is essential for its diverse energy metabolism.

A comprehensive proteomic approach was applied to identify candidate GSP proteins including those potentially involved in metal-reduction. Steady state $O_2$-limited wild type and *gspD* mutant cells of MR-1 were sampled from bioreactors for mass spectral proteomics and metal reduction activities. Proteome analysis of whole cells, cell fractions, membrane vesicles, and extracellular proteins revealed the mislocalization of several hypothetical proteins in the mutant compared to the wild type, as well as OmpW. OmpW and one hypothetical exhibited increased expression in cells incubated with various metals. The localization of MtrA, MtrB, and MtrC, proteins previously implicated in metal reduction, were unaffected according to proteome and western blot analyses. These results demonstrate that while the GSP is necessary for efficient metal reduction in these cells, several key electron transfer proteins essential for Fe(III) and Mn(IV) reduction were not mislocalized. By applying a combination of controlled cultivation integrated with proteome measurements genes that are candidate secreted proteins, including those involved in metal transformation, can be identified.

# 21

## Integrated Analysis of Gene Functions and Regulatory Networks Involved in Anaerobic Energy Metabolism of *Shewanella oneidensis* MR-1

**Jizhong Zhou**[1] (zhouj@ornl.gov), Dorothea K. Thompson[1], Matthew W. Fields[1], Timothy Palzkill[2], James M. Tiedje[3], Kenneth H. Nealson[4], Alex S. Beliaev[5], Ting Li[1], Xiufeng Wan[1], Steven Brown[1], Dawn Stanek[1], Weimin Gao[1], Feng Luo[1], Jianxin Zhong[1], Liyou Wu[1], Barua Soumitra[1], Crystal B. McAlvin[1], David Yang[1], Robert Hettich[1], Nathan VerBerkmoes[1], Yuri Gorby[5], Richard Smith[5], Mary Lipton[5], and James Cole[3]

[1]Oak Ridge National Laboratory, Oak Ridge, TN; [2]Baylor College of Medicine, Houston, TX; [3]Michigan State University, East Lansing, MI; [4]University of Southern California, Los Angeles, CA; and [5]Pacific Northwest National Laboratory, Richland, WA

Collaborating *Shewanella* Federation Team Leaders: Jim Fredrickson, Pacific Northwest National Laboratory, Richland, WA; Carol Giometti, Argonne National Laboratory, Argonne, IL; Eugene Kolker, BIATECH, Bothell, WA; and Monica Riley, Marine Biological Laboratory, Woods Hole, MA

*Shewanella oneidensis* MR-1, a facultatively anaerobic g-proteobacterium, possesses remarkably diverse respiratory capacities. In addition to utilizing oxygen as a terminal electron acceptor during aerobic respiration, *S. oneidensis* can anaerobically respire various organic and inorganic substrates, including fumarate, nitrate, nitrite, thiosulfate, elemental sulfur, trimethylamine *N*-oxide (TMAO), dimethyl sulfoxide (DMSO), Fe(III), Mn(III) and (IV), Cr(VI), and U(VI). However, the molecular mechanisms underlying the anaerobic respiratory versatility of MR-1, however, remain poorly understood. In this project, we have integrated genomic, proteomic and computational technologies to study energy metabolism of this bacterium from a systems-level perspective.

**Molecular Responses to Anaerobic Growth with Different Electron Acceptors**. To define the repertoire of MR-1 genes responding to different terminal electron acceptors, transcriptome profiles were examined in cells grown with fumarate, nitrate, thiosulfate, DMSO, TMAO, ferric citrate, ferric oxide, manganese dioxide, colloidal manganese, and cobalt using DNA microarrays covering ~99% of the total predicted protein-encoding open reading frames in *S. oneidensis*. Total RNA was isolated from cells grown anaerobically for 3.5 hours in the presence of different electron acceptors and compared to RNA extracted from cells grown under fumarate-reducing conditions (the reference condition). More than 1600 genes display significant expression changes across different metal-reducing conditions. Real-time PCR analysis for some selected genes showed that microarray-based quantitation is highly accurate. Hierarchical cluster analysis indicated that genes showing differential expression under metal-reducing conditions generally clustered together, whereas genes showing differences in mRNA abundance levels under non-metal respiratory conditions clustered together. Interestingly, no significant differences in expression profiles were observed between solid and soluble metal acceptors. Only a few genes specific for any particular metal were identified. In contrast, a group of over 150 genes displaying highly specific up-regulation under all metal-reducing conditions were identified, including, putative transporters, outer membrane components, as well as two electron transfer proteins (flavodoxin and a *c*-type cytochrome). In addition, a number of genes, of which 35-55% encoded hypothetical proteins, were uniquely induced or repressed in response to a single

electron acceptor. This work has yielded numerous candidates for targeted mutagenesis and represents an important step towards the goal of characterizing the anaerobic respiratory system of *S. oneidensis* MR-1 on a genomic scale.

**Phage Display**. Along with mass spectrometry, two-hybrid system and protein arrays, phage display is another powerful technique for studying protein-ligand interactions. The first key step of mapping protein interactions was to clone all protein-coding ORFs to allow exogeneous expression of its protein for functional analysis. We have spent tremendous efforts in cloning genes into an universal vector. Progress as of the close of 2003 is that 1,691 genes were cloned while no clones were obtained for 174 genes. Additionally, a random phage display library utilizing "shotgun" cloning of sheared *S. oneidensis* genomic DNA has been constructed.

**Genetic Mutagenesis**. One of the most powerful ways to define the function of a gene is to turn the gene off or change the expression by replacing the normal gene with a mutated counterpart. We have successfully modified and utilized vector systems for homologous recombination in *S. oneidensis* MR-1. Currently, our laboratory is interested in understanding transcriptional gene regulation in *S. oneidensis*. We are targeting approximately 220 annotated transcription factors (TFs) for knockout mutatgenesis. Analysis of the *S. oneidensis* genome sequence suggests that insertional mutagenesis is appropriate for only 78 of these TFs as they are transcribed in their own operon. We have also systematically knocked out some of the genes involved in metal reduction as revealed by microarray analysis discussed above.

Numerous other MR-1 genes have been successfully inactivated using a PCR-based, in-frame deletion mutagenesis strategy. Our current collection of deletion mutants includes those strains with mutations in *etrA*, *arcA*, *fur*, *crp*, *fur/etrA*, *etrA/crp*, *rpoH* (sigma-32), *ompR*, *envZ*, *oxyR*, *cya*1-3 (adenylate cyclases), and many others. Microarray-based gene expression profiling has been used to analyze a number of these mutant strains. For example, we have employed whole-genome DNA microarrays, large-scale proteomic analysis using liquid chromatography-mass spectrometry (LC-MS), and computational motif discovery tools to define the *S. oneidensis* Fur regulon. Using this integrated approach, we identified 9 probable operons (containing 24 genes) and 15 individual ORFs of either unknown function (SO0447-48-49, 0798-97, 0799, 1188-89-90, 2039, 3025, 3027, 3406-07-08, 3062, 3344, 4700, 4740) or annotated as encoding transport and binding proteins (*ftn*, *bfd*, SO1111-12, 1482, 1580, *feoAB*, *alcA*-3031-32, 3669-68-67, *tonB1-exbB1-exbD1*, *viuA*, *irgA*, 4743) that are predicted to be direct targets of Fur-mediated repression based on their up-regulated expression profiles in a *fur* deletion mutant and the presence of potential Fur-binding sites in their upstream regulatory regions. This study suggests, for the first time, a possible role of 4 operons and 8 ORFs of unknown function in iron metabolism.

**Chemostat Growth Studies with MR-1 Mutant Strains**. Using the growth facility at PNNL, *Shewanella oneidensis etrA* and *arcA* deletion strains and the parental strain were each grown in chemostats in continuous culture for 410 hours. The growth conditions were altered from an aerobic steady state, to a microoxic steady state and to an anaerobic steady state to examine the contribution of each regulator in *S. oneidensis*. Samples were collected at each steady state for organic acid, proteome, cytochrome and transcriptome analyses. Samples were also harvested at 0, 5, 10, 20, 30, 40, 50, 60, 90, 120, and 150 minutes after transition from aerobic to microoxic steady states for mRNA and protein analysis.

**Elucidation of the Functions of a Conserved Hypothetical Protein**. Whole-genome sequence analyses of a variety of microorganisms indicated that 30-60% of the identified genes encode functionally unknown proteins. Defining the functions of hypothetical proteins is a great challenge. Integrated approaches for systematic study of their functions are needed. As a first attempt, an in-frame deletion mutant was generated for the conserved hypothetical protein of 592 amino acids, SO1377. Physiological analysis showed that this mutant was very sensitive to hydrogen peroxide, showed slow growth rate under aerobic condition but not anaerobic conditions, and had higher spontaneous mutation rates. Microarray analysis revealed that numerous genes are affected by this mutation. Computational analyses of secondary and tertiary structure also revealed that the protein could have potential functions in formation of protein complexes at the inner bacterial cell membrane, ATP/GTP binding, nucleotide binding, protein transport and molecular chaperone. Overall, our results suggested this gene could be involved in iron homeostasis and oxidative damage protection in *S. oneidensis* MR-1.

**Molecular Basis of Stress Responses.** Other work related to Shewanella focuses on the elucidation of the molecular basis of bacterial adaptive responses to various environmental stresses, namely, heat stress, cold stress, high salt, low/high pH, oxidative stress, and metal toxicity. These studies employ primarily global gene expression profiling using cDNA/oligonucleotide microarrays and targeted gene mutagenesis. The initial manuscripts for two of these studies (heat shock and salt stress) have already been written and have been submitted for publication or are close to being submitted. In the study on oxidative stress, the effect of $H_2O_2$-induced oxidative stress on the gene expression profiles of *S. oneidensis* wild type and mutant strains was investigated. Microarray analysis of the wild type cells indicated significant changes in the expression levels of numerous genes that are known or have not been previously described to be involved in the oxidative stress responses of other bacterial species. Among these are the alkyl hydroperoxide reductase (*Ahp*) gene, the catalase (*Kat*) gene, the stress response DNA-binding protein (*dps*) gene, and the genes involved in the TonB transport systems. In addition, a LysR family transcriptional regulator showed immediate yet transient upregulation in response to $H_2O_2$ treatment, suggesting the hypothesis that it regulates $H_2O_2$ stress responses in *Shewanella oneidensis.* Sequence comparison and computational modeling predicted the gene to be the potential analog of the *E. coli* OxyR gene. Yet phenotype characterization of the deletion mutant of the gene revealed interesting responses toward various oxidative stimuli. Global expression profiling of the mutant indicated that the LysR regulator indeed controlled some of the genes that had been reported to belong to the OxyR regulon in *E. coli*, but it also regulated many uncharacterized genes.

In another study, we examined the response of *S. oneidensis* to high levels of heavy metals to better understand the repertoire of genes and regulatory mechanisms enabling heavy metal resistance. MR-1 was able to grow in LB medium with strontium ($Sr^{2+}$) concentrations as high as 180 mM, but showed substantial growth inhibition at levels above 180 mM. *S. oneidensis* resistance to 180 mM Sr was examined using DNA microarrays. Transcriptome profiles were generated from mid-exponential phase bacteria grown in the presence of $Sr^{2+}$ and compared to profiles from MR-1 cultured to the same growth phase in the absence of strontium. The stress response of *S. oneidensis* to a shock addition of 180 mM Sr was also examined after 5, 30, 60 and 90 minutes using microarrays. Siderophore biosynthesis and iron uptake genes were highly induced (up to 622 fold) and a siderophore biosynthetic mutant was more sensitive to strontium, suggesting that siderophore production plays an integral role in the ability of *S. oneidensis* to mediate strontium resistance.

**Network Modelling.** Understanding the regulatory interactions between thousands of genes in a given organism from massive time-course microarray data is one of the most challenging tasks in the field of microbial functional genomics. Currently, the inference of such genetic interaction networks is harmed by the dimensionality problem because the number of genes in a genome far exceeds the number of measured time points due to high cost of measurements. It is essential to develop powerful computational tools to extract as much biological information as possible from ambiguous expression data containing noise. Different from existing methods, we are developing a computational method based on random matrix theory. We are using the matrix of pair-wise correlation to identify connections between genes. In contrast to other network identification methods, the threshold for defining network links is determined automatically and self-consistently based on the data structure itself. We have applied this method to identify regulatory networks in yeast based on the massive available microarray data. The identified gene interactions were very consistent with our knowledge, suggesting that this method is very useful for network identification. We are now further testing this method based on microarray data from Shewanella, Deinococcus, yeast, worm, fly and human.

# 22

## Profiling *Shewanella oneidensis* Strain MR-1: Converting Hypothetical Genes into Real, Functional Proteins

**Eugene Kolker** (ekolker@biatech.org), Samuel Purvine, Alex F. Picone, Natali Kolker, and Tim Cherny

BIATECH, Bothell, WA

Collaborating ***Shewanella*** **Federation** Teams: **J. Fredrickson**, M. Romine, Y. Gorbi, A. Beliaev, B. Cannon (PNNL); **R. Smith, G. Anderson, K. Auberry, M. Lipton, D. Elias (PNNL); J. Tiedje**, X. Qiu, J. Cole (MSU); **K. Nealson**, S. Tsapin (USC); **M. Riley**, M. Serres (MBL); **C. Giometti**, G. Babnigg (ANL); **J. Zhou**, D. Thompson (ORNL).

Other **Collaborating** Teams: **E. Koonin**, M. Galperin, K. Makarova (NCBI); **C. Lawrence**, L.-A. McCue (WC); **B. Palsson**, A. Raghunathan, N. Price (UCSD); **H. Heffelfinger**, J. Timlin (SNL); **J. Yates**, W. Zhu (Scripps).

The progress in genome sequencing has led to a rapid accumulation in GenBank submissions of uncharacterized "hypothetical" proteins. These proteins, which have not been experimentally characterized and whose functions cannot be deduced from simple sequence comparisons alone, now comprise approximately one third of the public databases. That is, despite significant progress in the experimental research, this so called "70% hurdle" still holds, with every new genome bringing novel unknown proteins numbering in the hundreds or even thousands. Being very complex and fascinating in numerous aspects of its behavior and responses, *Shewanella oneidensis* strain MR-1 (SO) presents an even greater challenge, as over half of its predicted genes are considered hypothetical. If past performance in experimental characterization of new proteins from *Escherichia coli* K-12, roughly 25 per year, is of any predictive power, it will take many decades before the biological function of all these (SO) proteins is discovered.

Expression profiling of SO cells under multiple growth conditions done by the *Shewanella* Federation consortium was performed. Among the performed experiments are continuous and controlled batch cultures of SO cells under a variety of

different environmental conditions. These include electron acceptors and substrates, and limitations of thereof, such as $O_2$, $Ca^{2+}$, and Pi-limitations and UV-radiation stresses. Special emphasis was placed on robust, reproducible, and statistically validated results, rather than optimizing coverage of the expressed gene and protein contents for the above conditions. Earlier studies of SO presented a baseline of over 4,600 predicted genes with approximately 2,350 hypothetical ones.

SO gene profiling resulted in conservative estimation of over 4,000 expressed genes, including identification of over 1,900 hypothetical genes. Protein profiling experiments conservatively estimated approximately 1,550 expressed proteins with approximately 500 hypothetical ones. Using a combination of transcriptomic and proteomic approaches as well as statistical and computational methods, this analysis confidently identified over 450 hypothetical genes that were expressed in cells both as genes and proteins. In an attempt to understand the functions of these proteins, we used a variety of publicly available analysis tools. This resulted in exact or general functional assignments for over 200 hypothetical proteins. Accurate functional annotation of uncharacterized proteins calls for an integrative approach, combining expression studies with extensive computational analysis and curation, followed by the directed experimental verification.

# 23

## Systems Biology of *Shewanella oneidensis* MR-1: Physiology and Genomics of Nitrate Reduction, the Radiation Stress Response, and Bioinformatics Applications

**James M. Tiedje** (tiedjej@msu.edu), James R. Cole, Claribel Cruz-Garcia, Joel A. Klappenbach, and Xiaoyun Qiu

Michigan State University, East Lansing, MI

Collaborating *Shewanella Federation* Team Leaders: Jim Fredrickson, Margie Romine, Yuri Gorby (PNNL); Eugene Kolker (BIATECH); and Jizhong Zhou (ORNL)

**The Stress Response: Effects of Ultraviolet Radiation.** Successful application of *Shewanella oneidensis* MR-1 in bioremediation applications may necessitate cellular tolerance to toxic levels of pollutants and damage-inducing radiation. Solar ultraviolet radiation (UVR) is perhaps the most mutagenic agent to which many organisms are exposed due to its abundance. We systematically investigated the stress response in MR-1 following exposure to UVC, UVB and UVA radiation. MR-1 showed extremely high sensitivity to both far- and near-UV with a $D_{37}$ value (UVC) of 5.6% relative to *E. coli* K12. Photoreactivation conferred a significantly increased survival rate to MR-1 in both UVB and UVC irradiated cells: as much as 177- to 365-fold and 11- to 23-fold survival increase after UVC and UVB irradiation respectively. A significant UV mutability to rifampin resistance was detected in both UVC and UVB treated cells. Different gene expression profiles were observed after UVC, UVB and UVA treatments. More than 300 genes were up-regulated after UVA exposure whereas only about 100 genes were induced after UVC exposure. Although the SOS response occurred in all three treatments, the induction of key genes in the SOS regulon (e.g. *rec*A, *lex*A, *pol*B *etc*.) was most robust in response to UVC. Genes that are involved in protection from oxidative damage showed an increased expression level in both UVB and UVA treatments. Unexpectedly, we did not observe induction of genes encoding nucleotide excision repair (NER) compo-

nents (e.g. *uvr*A, *uvr*B and *uvr*D) in either UVB or UVC treatments. We were also unable to identify any potential SOS box upstream of *uvr*A, *uvr*B and *uvr*D. Complementation of *Pseudomonas aeruginosa* UA11079 (*uvr*A$^-$) with *uvr*A of MR-1 increased the UVC resistance of this strain more than three orders of magnitude, indicating the functionality of UvrA in repairing UVR-induced DNA damage. Using RT-PCR, we detected transcripts of *uvr*A, *uvr*B and *uvr*D from MR-1 in both UVR treated and untreated sample at equivalent levels, indicating that component genes of NER are constitutively expressed. Loss of the damage inducible NER system may contribute to the high sensitivity of this bacterium to UVR.

**Aerobic and Anaerobic Nitrate Reduction.** Nitrate is often found as a co-contaminant in metal and radionuclide contaminated groundwater, and understanding the response of *S. oneidensis* MR-1 to these compounds is critical to effective bioremediation applications. *S. oneidensis* MR-1 is capable of dissimilatory nitrate reduction to ammonia during both aerobic and anaerobic growth. Nitrate is reduced by *S. oneidensis* MR-1 in a stepwise manner from nitrate > nitrite > ammonia. Complete reduction of nitrate precedes initiation of nitrite reduction, a process controlled by thermodynamics. Genome analysis supports this physiology: *S. oneidensis* MR-1 possesses genes for a single nitrate reductase (NapA) and a single nitrite reductase (NfrA) that catalyze the reduction of nitrate to ammonia in two enzymatic steps. Expression of *napA* and *nrfA*, measured via quantitative PCR, is maximal in anaerobic batch cultures initiated with >0.5 mM nitrate. A decrease in *nap*A and *nrf*A gene expression with increasing concentrations of nitrate occurs in *E. coli*, indicating an alterative regulatory system is operating in *S. onedensis* MR-1. The expression of *nar*P/*nar*Q (NO3/NO2 sensor/response regulator) was constant in cultures fed up to 10 mM nitrate. During aerobic growth conditions with and without nitrate, *nap*A and *narP/narQ* were equivalently expressed, indicating constitutive expression of nitrate reductase activity independent of the presence oxygen or nitrate.

Nitrite accumulates during lactate-dependent nitrate reduction, and was found to inhibit growth in both aerobic and anaerobic batch culture. Decreased growth rates during chemostat culture did not alleviate nitrite toxicity due to the stepwise reduction of nitrate to ammonia. Anaerobic nitrate reduction was limited to the oxidation of lactate and pyruvate as sole carbon and energy sources - other sugars and carboxylic acids (and hydrogen) did not support growth. Growth limitation due to nitrite toxicity during aerobic batch culture was assessed using whole-genome microarrays. Significant up-regulation of genes encoding heat shock and DNA repair proteins that are associated with oxidative stress occurred in the presence of nitrite. Nitrite also resulted in the significant down-regulation of many genes involved in iron acquisition, possibly as a mechanism of reducing DNA damage induced by hydroxyl radicals generated via intracellular iron oxidation. The capacity for nitrate reduction of *S. oneidensis* MR-1 may therefore be limited by its ability to mediate oxidative damage induced by nitrite accumulation.

***MicroPlateDB* – a LIMS for Quality Control and Data Archiving Microplate Data.** In a multi-investigator research effort, such as the *Shewanella Federation*, open-access to research data and protocols is critical to genomics-level investigations involving tools such as microarrays and proteomics. We have continued development of an internet-browser accessible laboratory information management system (LIMS), the '*MicroPlateDB*', for tracking and archiving data generated during microarray construction. The LIMS is structured with the laboratory microplate as the central data type and the contents of plates are combined during virtual "reactions" as they are carried out in the. Customization of the LIMS is controlled by a set of basic data tables containing information on microplate types, contents, and

how contents of microplates are combined and stored during laboratory procedures. User-level permissions control LIMS access and allow a project manager to specify the ability to view and/or modify data on an individual basis. With a login name and password, an investigator performing microarray studies can access the LIMS to find a gene of interest in the plate used to print the array, including the concentration, size, and gel-resolved quality of the PCR-product. The user also has the ability to 'drill-down' through a set of hyperlinked microplate graphic representations to track a PCR-product from a spot on the microarray to the primers and template used to create that product. Enhancements currently under development include user-defined searching and an open-source version for public release. The LIMS was initially customized to track process information obtained during the production of a PCR-based DNA microarray for *S. oneidensis* MR-1 and has also been chosen for use in several other microarray construction projects, including the ORNL *Deinococcus radiodurans* microarray.

# 24

## Development and Application of Optical Methods for Characterization of Protein-Protein Interactions in *Shewanella oneidensis* MR-1

Natalie R. Gassman[1*] (ngassman@chem.ucla.edu), Achillefs N. Kapanidis[1], Nam Ki Lee[1, 4], Ted A. Laurence[1,5], Xiangxu Kong[1], and **Shimon Weiss**[1,2,3]

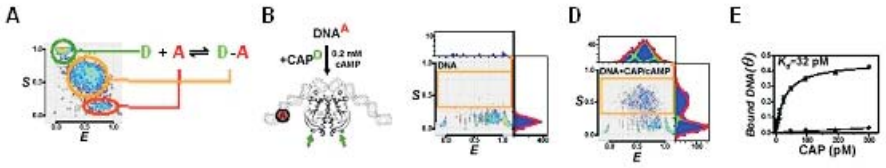
*Presenting author

[1]Dept. of Chemistry and Biochemistry, University of California, Los Angeles, CA; [2]Dept. of Physiology, David Geffen School of Medicine, University of California, Los Angeles, CA; [3]California NanoSystems Institute, University of California, Los Angeles and Santa Barbara, CA; [4]Seoul National University, Seoul, Korea; and [5]Lawrence Livermore National Laboratory, Livermore, CA

The increased availability of microbial genomes has increased the drive to elucidate the complex biological networks these microbes utilize to adapt to extreme environmental conditions. Regulation, cell adhesion, and respiration networks, which accomplish bioremediation of metals and radionuclides, are of particular interest to the Genomics:GTL program. *Shewanella oneidensis* MR-1, a dissimiliatory metal reducing bacterium with the ability to utilize a large variety of electron acceptors, is ideal for bioremediation applications. While widely studied, the mechanism by which MR-1 utilizes these electron acceptors remains unclear. Understanding these complex respiration mechanisms requires the characterization of post-translational macromolecular interactions. We are investigating these numerous protein-protein

interactions in MR-1 by developing novel optical methods for their detection and characterization.

Figure 1. Analysis of a protein-DNA interaction using ALEX-FAMS. A. E-S histogram for D-only, A-only, and D-A species with different RD-A. B. Model of CAP-DNA complex and labeling scheme. Acceptor (red) was placed on DNA, and the donor was placed on 2 possible sites on CAP (green arrows). C. E-S histogram of A-containing species, DNAA. D. E-S histogram of DNAA incubated with CAPd and 0.2 mM cAMP. E. Single molecule sorting allows calculation of biomolecular constants; ALEX-based titration of DNA with CAP in the presence (filled circles) or absence of cAMP (open circles).




We have expanded on a current single-molecule fluorescence spectroscopy (SMFS) method, single-pair Förster resonance energy transfer (spFRET), to measure stochiometry and interactions. spFRET uses a single laser to probe the transfer of excitation energy from a donor (D) fluorophore to a complementary acceptor (A) fluorophore of an interacting pair, yielding a D-A distance sensitive value *E*, which acts as a "spectroscopic ruler" for the 1-10 nm scale. While an excellent qualitative indicator of molecular interactions, the limited dynamic range of the FRET ruler precludes the measurement of interactions between large macromolecules and/or multimeric complexes. By using an alternating-laser excitation (ALEX) scheme, we have expanded the spFRET technique to report on structure, dynamics, stoichiometries, local environment and molecular interactions. This is accomplished by obtaining D-excitation and A-excitation–based observables for *ach single molecule* by rapidly alternating between D-excitation and A-excitation lasers. This scheme probes directly both FRET donors and acceptors present in a single diffusing complex and recovers distinct emission signatures for all species involved in interactions by calculating two fluorescence ratios: the FRET efficiency *E*, a distance-based ratio which reports on conformational status of the species, and a new, distance-independent stoichiometry-based ratio, *S*, which reports on the association status of the species. Two-dimensional histograms of E and *S* allow virtual sorting of single molecules by conformation and association status (Fig. 1A). ALEX is a homogeneous, "mix-and-read" assay, where interacting species are combined and optical readouts report *simultaneously* on their association status and conformational status. The potential applications of this methodology are extensive and characterization of known protein-DNA interactions, *Escherichia coli* catabolite activator protein (CAP) with DNA (Fig. 1B-D), has illustrated the method's robust nature.

The complex regulatory mechanisms governing the expression of genes involved in electron transport and energy generation in MR-1 provide a diverse array of protein-DNA and protein-protein interactions that are ideally suited for the ALEX method. One such regulatory mechanism, activated under environmental stress, is the two-component signaling cascade that initiates gene expression by the alternative sigma factor, $\sigma^{54}$. Transcriptional regulation is achieved through a cascade of protein-protein interaction that results in the interaction of a transcription regulator with the $\sigma^{54}$-RNA polymerase (RNAP) holoenzyme complex to initiate transcription. One example of this signaling cascade is the interaction of a nitrogen regulatory protein (NtrC) with $\sigma^{54}$-RNAP holoenzyme to initiate transcription of genes involved in nitrogen fixation in MR-1. Upon stimulus by environmental stress, a

sensor protein autophosphorylates resulting in the downstream phosphorylation the transcriptional regulator, NtrC. An NtrC oligomeric form then binds upstream of the promoter region and via a looped DNA intermediate catalyzes the formation of the open transcription complex. Using the ALEX methodology, we can now examine the mechanistic process of gene regulation under stress conditions from the oligomerization of the transcription regulator to the activating interaction between NtrC and the $\sigma^{54}$-RNAP holoenzyme to initiation of transcription. Progress in protein expression, site-directed mutagenesis and fluorescence labeling of MR-1 NtrC, $\sigma^{54}$-and RNAP holoenzyme will be reported.

# 25

## Annotation of Genes and Metabolism of *Shewanella oneidensis* MR-1

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

Marine Biological Laboratory, Woods Hole, MA

### Annotation

Our continuing annotation of the genes and gene products of *Shewanella oneidensis* MR-1 is taking advantage of two seldom used sources of information on protein function: (1) functions of structural domains within protein sequences, and (2) the functions of paralogous groups, groups of genes that seem to have descended from the same ancestor and tend to retain related functions.

The Structural Classification of Proteins database (SCOP) has a section describing Superfamilies of structural domains. Using a HMM method, the presence and location of structural domains has been determined for some genomes. The data for *S. oneidensis* finds structural domains in 2570 of the proteins. The data has been scrutinized in particular for all open reading frames (ORFs) having no functional assignment. 366 of unknown ORFs could be assigned some information on function from this connection.

Many bacterial genomes have genes for proteins that appear to have arisen during evolution by duplication followed by divergence. Distantly but firmly related proteins have been assembled by collecting related sequences (determined by the Darwin algorithm) into groups by transitive relationships. Such groups of sequence similar proteins can include only distantly related members in that similarity to only one protein in the group is sufficient for inclusion. No protein is a member of more than one group. Using this approach, 408 paralogous groups were identified, ranging in size from 2 to 64 members per group.

Paralogous groups give some insight to the numbers of ancestral genes required to generate contemporary bacterial gene families. Also, since paralogous group member of known function show similarity of function (sometimes closely related, sometimes more distantly), any unknown members of a paralogous group can be assigned the common denominator of function for that group.

### Metabolism

To survey the metabolic capabilities or *S. oneidensis*, we recorded similarities to the protein sequences of 50 fully sequenced organisms, again using the Darwin algo-

rithm. The organisms with the largest number of "hits" were *Vibrio cholera*, *Yersinia tuberculosis*, *Escherichia coli* and *Pseudomonas aeruginosa*. Of these, biochemical information for gene products is far and away greatest for *E. coli*. Thus it was possible to assign putative enzyme function and pathway existence when homologs exist in *E. coli*, but not possible when there was similarity to a protein of unknown function in one of the other bacteria, no analogous gene in *E. coli*. Therefore broadly speaking, the metabolic capacities of *S. oneidensis* are similar to those of *E. coli*, but this is partly because more *E. coli* proteins have been characterized than in the other bacteria. Nevertheless, there were a few functions in for instance *Pseudomonas aeruginosa* such as part of the beta-ketoadipate pathway that seem to be present in *S. oneidensis* but not in *E. coli*. With further exploration of the biochemistry of organisms other than *E. coli*, more assignment of biochemical capability will be possible.

Biosynthetic capabilities for small molecule cofactors, carriers are largely intact. However in most cases where *E. coli* has two or more isozymes, *S. oneidensis* has only one. At present the broad picture for utilization of carbon sources involves very few 5 or 6 carbon sugars and sugar derivatives, rather evidence for the utilization of 3 or 2 carbon carbohydrates and organic acids. There is a defect in an essential enzyme of the glycolytic pathway. However the enzymes for utilization of some 6 carbon sugars, e.g. galactose are present and the Entner-Douderoff pathway seems to be present, completely adequate to serve 5 and 6 carbon substrates. Therefore it is not clear why many 5 and 6 carbon compounds are not utilized. This aspect bears experimental exploration.

Only some of the enzymes for reduction of organic terminal electron acceptors that are found in *E. coli* seem to be present in *S. oneidensis*. Formate metabolism is present. The well-known use of metal ions as electron acceptors could abrogate the need for using many organic acceptors. Many electron transfer intermediates are present, consistent with the unusual richness of energy transfer by this organism.

## Institute for Biological Energy Alternatives

# 26

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

John I. Glass[1], Nina Alperovich[1], Nacyra Assad-Garcia[1], Holly Baden-Tillson[1], Hoda Khouri[2], Matt Lewis[3], William C. Nierman[2], William C. Nelson[2], Cynthia Pfannkoch[1], Karin Remington[1], Shibu Yooseph[1], Hamilton O. Smith[1], and **J. Craig Venter**[1] (jcventer@tcag.org)

[1]Institute for Biological Energy Alternatives, Rockville, Maryland; [2]The Institute for Genomic Research, Rockville, MD; and [3]The J. Craig Venter Science Foundation Joint Technology Center, Rockville, MD

IBEA 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 are using two mycoplasma species as platforms to develop methods for construction of a minimal cell. *Mycoplasma genitalium* is a slow-growing human urogenital pathogen that has the smallest known genome of any free-living cell at 580 kb. It has already been used to make a preliminary estimate of the minimal gene set. Global transposon mutagenesis identified 130 of the 480 *M. genitalium* protein-coding genes not essential for cell growth under laboratory conditions. That study also predicted there may be as many as 85 other *M. genitalium* genes that are similarly not essential. *Mycoplasma capricolum* subsp. *Capricolum*, an organism endemic in goats, was chosen as another platform because of its rapid growth rate and reported genetic malleability. To facilitate work with this species we sequenced and annotated its 1,010,023 bp genome. In anticipation of eventually synthesizing artificial genomes containing a minimal set of genes necessary to sustain a viable replicating bacterial cell we took two approaches to determine the composition of that gene set.

In one approach we used global transposon mutagenesis to identify non-essential genes in both of our two platform mycoplasma species. We created, isolated, and expanded clonal populations of sets of random mutants. Transposon insertion sites were determined by sequencing directly from mycoplasma genomic DNA. This effort has already expanded the previously determined list of non-essential *M. genitalium* genes, and in this study, because we isolated and propagated each mutant, we can characterize the phenotypic effects of the mutations on growth rate and colony morphology. Additionally, identification of non-essential genes in our two distantly related mycoplasma species permits a better estimate of the essential mycoplasma gene set.

In our other approach, we analyzed 11 complete and 3 partially sequenced mycoplasma genomes to define a consensus mycoplasma gene set. 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 less than 100 genes common to all bacteria; however determination of conserved genes within the narrow mycoplasma taxon is much more instructive. The combination of comparative genomics with 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 parologous gene families likely involved in antigenic variation that comprise significant fractions of each genome and presumably unnecessary for cell viability under laboratory conditions.

# 27

## Whole Genome Assembly of Infectious ϕX174 Bacteriophage from Synthetic Oligonucleotides.

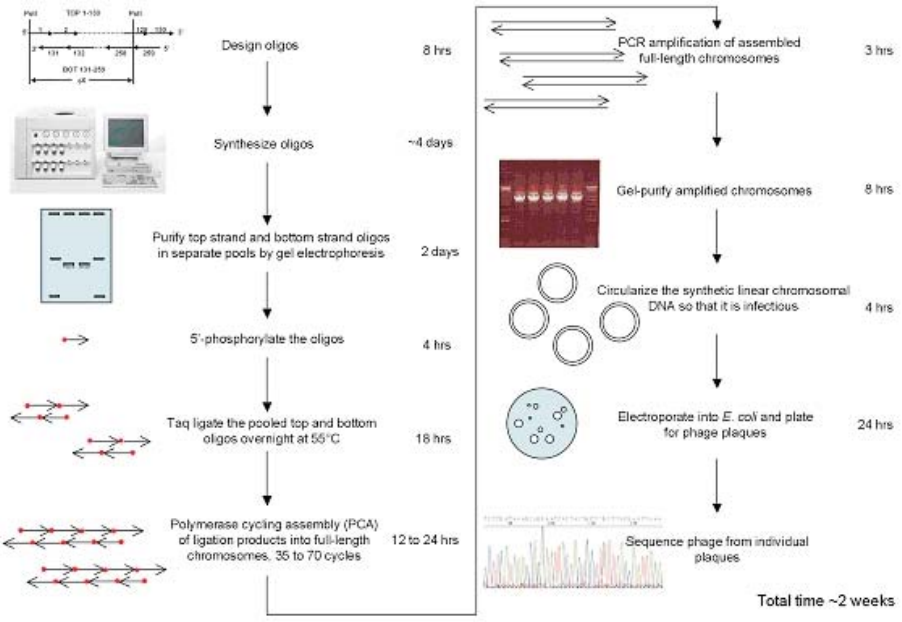
Hamilton O. Smith[1], Clyde A. Hutchison III[2], Cynthia Pfannkoch[1], and **J. Craig Venter**[1] (jcventer@tcag.org)

[1]Institute for Biological Energy Alternatives, Rockville, MD and [2]Department of Microbiology and Immunology, University of North Carolina, Chapel Hill, NC

We have improved upon the methodology and dramatically shortened the time required for accurate assembly of 5 to 6 kb segments of DNA from synthetic oligonucleotides. As a test of this methodology we have established conditions for the rapid (14 days) assembly of the complete infectious genome of bacteriophage ϕX174 (5,386 bp) from a single pool of chemically synthesized oligonucleotides. The procedure involves three key steps: 1) Gel purification of pooled oligonucleotides to reduce contamination with molecules of incorrect chain length, 2) Ligation of the oligonucleotides under stringent annealing conditions (55C) to select against annealing of molecules with incorrect sequences, and 3) Assembly of ligation products into full length genomes by polymerase cycling assembly (PCA), a non-exponential reaction in which each terminal oligonucleotide can be extended only once to produce a full-length molecule. We observed a discrete band of full-length assemblies upon gel analysis of the PCA product, without any PCR amplification. PCR amplification was then used to obtain larger amounts of pure full-length genomes for circularization and infectivity measurements. The synthetic DNA had a lower infectivity than natural DNA, indicating approximately one lethal error per 500 bp. However, fully infectious ϕX174 virions were recovered following electroporation into *E. coli*. Sequence analysis of several infectious isolates verified the accuracy of these synthetic genomes. One such isolate had exactly the intended sequence. We propose to assemble larger genomes by joining separately assembled 5 to 6 kb segments; approximately 60 such segments would be required for a minimal cellular genome. Below is a schematic diagram of the steps in the global synthesis of infectious ϕX174 bacteriophage from synthetic oligonucleotides.

The power of the above global assembly method will be fully realized when methods to remove errors from the final product are developed. Further experiments are underway to increase the efficiency of error correction.

Fig. 1. Schematic diagram of the steps in the global synthesis of infectious ϕX174 bacteriophage from synthetic oligonucleotides.




# 28

## Development of a *Deinococcus radiodurans* Homologous Recombination System

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

Institute for Biological Energy Alternatives, 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.

As a case in point, there are two candidates for the single-stranded DNA binding protein, Ssb (Dr0099 and Dr0100). To determine which of the two is the real Ssb, we first resequenced the Ssb region. We discovered two single-base deletions that when corrected give rise to a contiguous gene that contains two Ssb OB fold domains. We have purified the recombinant protein almost to homogeneity and characterized its DNA binding and strand-exchange properties. Our results suggest that despite some minor differences, the *Deinococcus* Ssb is very similar to the *E. coli* protein. In addition, using antibodies we have raised against DrSsb, we have determined that the amount of DrSsb protein, like recA, increases in the cell when exposed to a DNA damaging agent.

# 29

## Environmental Genome Shotgun Sequencing of the Sargasso Sea

**J. Craig Venter** (jcventer@tcag.org), Karin Remington, Jeff Hoffman, Holly Baden-Tillson, Cynthia Pfannkoch, and Hamilton O. Smith

Institute for Biological Energy Alternatives, Rockville, MD

We have applied whole genome shotgun sequencing to pooled environmental DNA samples in this study to test whether new genomic approaches can be effectively applied to gene and species discovery and to overall environmental characterization. To help ensure a tractable pilot study, we sampled in the Sargasso Sea, a nutrient-limited, open ocean environment. Further, we concentrated on the genetic material captured on filters sized to isolate primarily microbial inhabitants of the environment, leaving detailed analysis of dissolved DNA and viral particles on one end of the size spectrum, and eukaryotic inhabitants on the other, for subsequent studies.

Surface water samples were collected from three sites off the coast of Bermuda in February 2003. Additional samples were collected from a neighboring fourth site in May 2003. Genomic DNA was extracted from filters of 0.1 to 3.0 microns, and genomic libraries with insert sizes ranging from 2-6kb were made and sequenced from both ends. The 1.66 million sequences from the February samples were pooled and assembled to provide a single master assembly for comparative purposes. An additional 325,608 reads from the May samples were also analyzed. The assembly generated 64,398 scaffolds ranging in size from 826 bp to 2.1 Mbp, containing 256 Mbp of unique sequence and spanning 400 Mbp. Evidence-based gene finding revealed 1,214,207 genes within this dataset, including 1412 distinct small subunit rRNA genes. With this set of rRNA genes, using a 97% sequence similarity cut-off to distinguish unique phylotypes, we identified 148 novel phylotypes in our

sample when compared against the RDP II database[2]. Because the copy number of rRNA genes varies greatly between taxa (more than an order of magnitude among prokaryotes), rRNA-based phylogeny studies can be misleading. Therefore, we constructed phylogenetic trees using various other represented phylogenetic markers found in our dataset. Assignment to phylogenetic groups shows a broad consensus among the different phylogenetic markers.

Just as phylogenetic classification is strengthened by a more comprehensive marker set, so too is the estimation of species richness. In this analysis, we define "genomic" species as a clustering of assemblies or unassembled reads more than 94% identical on the nucleotide level. This cut-off, adjusted for the protein-coding marker genes, is roughly comparable to the 97% cut-off traditionally used for rRNA. Thus-defined, the mean number of species at the point of deepest coverage was 451; this serves as the most conservative estimate of species richness. However, in most of the samples we observed an average maximum abundance of only 3.3%. This is a level of diversity akin to what has been observed in terrestrial samples[3].

While counts of observed species in a sample are directly obtainable, the true number of distinct species within a sample is almost certainly greater than that which can be observed by finite sequence sampling. Modeling based on assembly depth of coverage indicates that there are at least 1,800 species in the combined sample, and that a minimum of 12-fold deeper sampling would be required to obtain 95% of the unique sequence. Further, the depth of coverage modeling is consistent with as much as 80% of the assembled sequence being contributed by organisms at very low individual abundance, compatible with total diversity orders of magnitude greater than the lower bound just given. The assembly coverage data also implies that more than 100Mbp of genome (i.e., probably more than 50 species) is present at coverage high enough to permit assembly of a complete or nearly-complete genome were we to sequence to 5- to 10-fold greater sampling depth.

We demonstrate the utility of such a dataset with a study of genes relevant to photobiology within the Sargasso Sea. The recent discovery of a homolog of bacteriorhodopsin in an uncultured γ-proteobacteria from the Monterey Bay revealed the basis of a novel form of phototrophy in marine systems[4] that was observed previously by oceanographers[5,6]. Environmental culture-independent gene surveys with PCR, have since shown that proteorhodopsin is not limited to a single oceanographic location, and revealed some 67 additional closely related proteorhodopsin homologs[7]. More than 782 rhodopsin homologs were identified within our dataset, increasing the total number of identified proteorhodopsins by almost an order of magnitude. In total, we have identified 13 distinct subfamilies of rhodopsin-like genes. These include four families of proteins known from cultured organisms (halorhodpsin, bacterorhodopsin, sensory opsins, and fungal opsin), and 9 families from uncultured species of which 7 are only known from the Sargasso Sea populations.

While we are a long way from a full understanding of the biology of the organisms sampled here, even this relatively small study demonstrates areas where important insights may be gained from the comprehensive nature of this approach. Our assembly results demonstrate one can apply whole-genome assembly algorithms successfully in an environmental context, with the only real limitation being the sequencing cost.

## References


1. The authors acknowledge the significant contributions of their collaborators on this project: J. Heidelberg, J.A. Eisen, D. Wu, I. Paulsen, K.E. Nelson, W. Nelson, D. E. Fouts, O. White and J. Peterson at The Institute for Genomic Research, A.L. Halpern, D. Rusch, and S.l Levy at The Center for the Advancement of Genomics, A. H. Knap, M. W. Lomas and R. Parsons at the Bermuda Biological Station for Research, Y. Rogers at the JCVSF Joint Technology Center, and K. Nealson at the University of Southern California.

2. J. R. Cole et al., *Nucleic Acids Research* **31**, 442 (Jan 1, 2003).

3. T. P. Curtis, W. T. Sloan, J. W. Scannell, *Proceedings of the National Academy of Sciences of the United States of America* **99**, 10494 (AUG 6, 2002).

4. O. Beja et al., *Science* **289**, 1902 (Sep 15, 2000).

5. Z. S. Kolber, C. L. Van Dover, R. A. Niederman, P. G. Falkowski, *Nature* **407**, 177 (Sep 14, 2000).

6. Z. S. Kolber et al., *Science* **292**, 2492 (Jun 29, 2001).

7. G. Sabehi et al., *Environ Microbiol* **5**, 842 (Oct, 2003).