Addendum

Abstracts received after January 24, 2005

Metagenome Analysis of Contaminated Sediments at the DOE Hanford Site

Natalia Maltsev[1], Tanuja Bompada[1], Banu Gopalan[2] (agor@ornl.gov), Shu-mei Li[3], Weiwen Zhang[3], J. Chris Detter[4], Paul Richardson[4], Margie Romine[3], and **Fred Brockman[3]**

[1]Bioinformatics Group, Argonne National Laboratory; [2]Computational Biology and Bioinformatics, Pacific Northwest National Laboratory, Richland, WA; [3]Microbiology Group, Pacific Northwest National Laboratory, Richland, WA; [4]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA

Technologies are needed to make improved inferences of microbial community function from metagenome sequence. Most microbes have evolved multiple mechanisms (programs) to capture energy. The cells' immediate biochemical and geochemical environment determines the regulatory networks that are engaged to run the best program to capture energy. Bulk extraction of RNA and analysis on microarrays largely destroys the spatial and functional linkages that are the key to understanding how communities interact. Therefore, a critical need for understanding the details of how multi-species microbial communities interact at the cellular level to generate a functional output, is the ability to interrogate the microscopic spatial organization of gene expression in these multi-species microbial communities.

The most likely methods to accomplish this objective are (1) robotic single (prokaryotic) cell "picking" by laser catapulting microscopy followed by RNA extraction, whole transcriptome amplification (if required), and probing and/or sequencing; or (2) mRNA-targeted non-PCR based fluorescence in situ hybridization (mRNA-FISH).

A second critical technology is the use of high-end computing to utilize massive amounts of metagenome sequence to design optimal phylogenetically-constrained function-oriented oligonucleotide probes for both of these approaches.

The goals of this newly funded project are to:

- Develop mRNA-targeted non-PCR based fluorescence in situ hybridization (mRNA-FISH) using soluble and nanoparticle near-infrared (low noise) probes coupled to advanced microscopy able to detect a very small number of photons
- Use grid-based computational tools to analyze community metagenome data to develop hypotheses regarding the functional processes and linkages occurring in the multi-species community, and for design of a suite of phylogenetically-constrained metabolic function "signature" probes.

The results presented in this poster relate to the second goal. As an initial exercise, we are analyzing metagenome sequence produced in a previous Microbial Genome Project. That project pooled enrichments from contaminated sediments beneath a leaking high-level radioactive waste tank at the DOE Hanford Site. Because biomass levels were very low (~10^4 cells per gram), a variety of enrichments were pooled in order to have adequate DNA to construct a clone library for sequencing. Most of the enrichments produced so little biomass that DNA concentrations were inadequate (in 2002) for constructing clone libraries. In 2003, a clone library was made from an enrichment pool that had the highest amount of DNA (only 750 nanograms). Community DNA's were also extracted in 2002 from pools of enrichments derived from more contaminated sediments, and in light of recent technological advancements clone libraries are now being constructed from those DNA samples (5 - 50 nanogram amounts) by Lucigen Corporation (Madison, WI).

Here we report on the preliminary metagenome analysis of the clone library constructed in 2003. Although the initial analysis reported here is a very small amount of sequence, we plan a minimum of 100-fold higher amounts of metagenome sequence from each library currently under construction. A total of 2,887 bacterial clones were sequenced and yielded a total of 7,071 hits representing 489 EC classes and 113 KEGG maps. At least one gene was present for synthesis of 18 of the 20 amino acids. Pathways in which ten or more genes in the pathway were present include metabolism of purines, pyrimidines, aminoacyl-tRNA, glycolysis/gluconeogenesis, pyruvate, starch, glycerolipid, porphyrin, glycine/serine/threonine, arginine/proline; and phenylalanine/tyrosine/tryptophan.

Protein hits were largely consistent with amplified and sequenced 16S rDNA phylogenies from both pooled enrichments and sediments. The 16S data from pooled enrichments showed 10 genera from the Micrococcineae, Propionibacterineae, and Steptomycineae suborders within the Actinobacteria (high GC Gram positive) phylum; and one genera (*Pseudomonas*) within the gamma class of the Proteobacteria phylum). The protein hits were 68% Proteobacteria, 30% Actinobacteria, and 3% to the Bacilli and Clostridia classes of the Firmicutes (low GC Gram positive) phylum. The *Pseudomonas* species detected in the DNA and protein hits are nitrate-reducers and are rare or absent in pristine, deep subsurface sediments at the Hanford Site; however, their presence is consistent with nitrate being the predominant inorganic contaminant in the sediments.

A web site has been constructed displaying taxonomic analysis of the metagenome; views of each contig including CDS information, potential functions, and relevant metabolic pathways; metagenome metabolic reconstruction; metabolic pathways indexed by similarity to organisms; and list of KEGG map identifications linked to organisms. Visualizations of the data using PNNL-developed Biological Data Fusion and OmniViz software will be shown. Interesting findings include (1) quite low and amino acid identities (and e-scores) for hits to the Actinobacteria and Firmicutes, suggesting a relatively novel component of this metagenome in comparison to current microbial and metagenome sequence from these phyla, and (2) a strong over-representation of transmembrane transport protein hits in the metagenome sequence. The relevance of these findings is being analyzed in detail. Future work includes metagenome sequence analysis of the very highly contaminated Hanford sediments and identification of phylogenetically-constrained metabolic function "signature" probes.

SAXS/WAXS Studies of σ54-Dependent AAA+ ATPases: Insights about Signal Transduction and Motor Function.

B. Tracy Nixon[1] (btn1@psu.edu), Baoyu Chen,[1] Michaeleen Doucleff,[2] David E. Wemmer,[2,3] Timothy R. Hoover,[4] and Elena Kondrashkina.[5]

1) Biochemistry and Molecular Biology, Penn State University, University Park, PA 16802 USA; 2) Chemistry, University of California, Berkeley, CA USA; 3) Lawrence Berkeley National Lab, Berkeley, CA USA; 4) Microbiology, University of Georgia, Athens, GA USA; 5) BioCAT at APS/Argonne National Lab, Illinois Institute of Technology, 9700 South Cass Ave, Argonne, IL 60439, USA.

AAA+ ATPases are molecular motors that provide important biological functions in all kingdoms of life. We are still learning how their actions are controlled and how they perform mechanical work. The most prevalent information processing system in bacteria, two-component signal transduction, is sometimes used to regulate the assembly AAA+ ATPase machines that regulate transcription by the σ54-form of RNA polymerase. Powerful and complementary microscope technologies are being developed under the GTL project ‘Microscopes of Molecular Machines (M3): Structural Dynamics of Gene Regulations in Bacteria’ (C*arlos Bustamante*, PI) to study the structure and changes that occur when multi-protein molecular machines such as AAA+ ATPases are formed. The technologies (cryo-electron microscopy, atomic force microscopy, optical tweezers, and single-molecule fluorescence microscopy) look at molecular machines from different but complementary perspectives. Cryo-electron microscopy can, for example, form visual images of an entire molecular complex, while single-molecule fluorescence can show real time formation of complexes as a result of fluorescence signals that can be seen as specific proteins come in contact with each other. Although static light scattering from molecules in solution occurs from molecules in all possible orientations, shape information is available to about 5 Angstroms resolution. The BioCAT beamline 18ID of the Advance Photon Source in Chicago is well suited for collecting such data, especially for large molecules for which very low Q data are desired. The research presented in this poster demonstrates how solution structures derived *ab initio* from small- and wide-angle X-ray scattering (SAXS/WAXS) data complement the microscopy approaches (comparable cryo-electron microscopy data will be presented separately by Sacha de Carlo and Eva Nogales).

We have collected scattering data for several proteins or protein fragments of DctD, NtrC, NtrC1 and PspF proteins, four such AAA+ ATPases. Solution structures determined from the scattering data give us insight into regulation and function of these molecular motors. In one case, a regulatory domain adopts two homo-dimeric forms, alternately repressing or derepressing motor assembly by adjacent ATPase domains; in another case, regulatory and ATPase domains cooperate to stabilize the assembled motor. Structures of ATPase in the presence of nucleotide analogs promise to reveal subdomain reorientations that are coupled with conformational changes in the ‘second region of homology’ and pore region of the ring shaped motors to mediate interaction with the target protein, σ54. Scattering data also yield preliminary models to explain how σ54 binds tightly to the activator in the transition state for ATP hydrolysis.

Use of the Advanced Photon Source was supported by the Department of Energy under contract W-31-109-ENG-38. BioCAT is a Research Center supported by National Institutes of Health, grant RR08630.

Cell-Free Protein Synthesis for High-Through-Put Proteomics
MacConnell Research Corporporation

Evan Dushman, Randal Sivila, and Jennifer Holmes and William P. MacConnell

The production of proteins from cloned DNA sequences is an important process for functional genomic studies and structural analysis, as well as many research applications including pharmaceutical drug discovery. We are developing new methodology and products for cell-free protein synthesis that allow production of up to 100 milligrams of protein using an inexpensive and highly stable wheat germ cell-free system. This system offers a tremendous advantage in simplicity and cost over *in vivo* protein expression methods in that it simplifies or eliminates: vector construction, cell transfection, and uncertainties of host cell synthesis. The method also allows the expression of multiple proteins in parallel, and can begin with PCR-generated DNA templates.

The overall objective of this work is to develop affordable *in vitro* protein synthesis reagents that will produce up to 100 milligrams of highly active protein using a universally applicable protocol. A simple processing instrument is also being developed to automate the protein synthesis reaction steps.

Results thus far demonstrate that our enhanced S30 wheat germ lysate can generate up to 30 milligrams of enzymatically active protein in one reaction. The process begins with double-stranded template DNA that is transcribed into mRNA (non-capped) using T7 RNA polymerase. A typical mRNA transcript is designed to contain the coding domain of the desired protein downstream from a strong ribosome binding sequence such as the TMV 5'UTR. The synthesized protein can be produced from either plasmid or PCR generated template DNA. In the case of PCR template, we begin with genomic DNA that was amplified with a first set of primers, then re-amplified with subsequent primer sets to add the T7, UTR and/or affinity tag sequences to the message or protein.

We have used the system to synthesize seven different proteins of varying sizes from bacterial and mammalian origin. Two of these proteins were successfully purified from the reaction mixture using a 6-his tag affinity purification method. The wheat germ system has been shown to be scaleable in trials where the energy generating reagents were added sequentially or when these components were diffused into the reaction through a permeable membrane. The in vitro method allows for the expression of toxic proteins that are impossible to produce by cellular expression methods. The system can also be used to generate protein from a predicted, but unknown, coding sequence or multiple variations of the same protein.

Several products will arise from this technology that can be sold directly by our company to laboratories throughout the world that perform genomic and proteomic research. We estimate that purified protein can be synthesized by this system for $13 per milligram. These products will save time and labor, improve the outcome of experiments and reduce the cost of small-scale protein production.

DOE Grant Number: DE-FG02-03ER83706

Decipherable Principles of Gene Regulation are Decipherable with Minimal Knowledge*

Dat H. Nguyen,[†] Patrik D'haeseleer, George M. Church

Department of Genetics
Harvard Medical School
77 Avenue Louis Pasteur
Boston, MA 02115

Gene regulation is responsible for organismal complexity and diversity in the course of biological evolution and adaptation, and it is determined primarily by the context-dependent behaviors of cis-regulatory elements (CRE's). Therefore, determining principles underlying their behaviors constitutes a fundamental objective of quantitative biology, yet this remains poorly understood. One major obstacle has been the lack of a good mathematical strategy for deciphering principles of gene regulation at a fine enough level of detail in order to distinguish the intricacy of regulatory signals encoded in the genome. Here we present a deterministic mathematical strategy, the Nguyen-D'haeseleer-Church (NDC) method, for deriving gene regulation principles in eukaryotic genomes. Unlike any other method, NDC works on all genes without assumption about regulation rules, or gene cluster membership, or manual tuning of parameters, while providing flexible and natural framework for incorporating experimental data such as ChIP-based protein-DNA interactions data. We will discuss many classes of gene regulation principles that CRE's obey in order to control gene expression we discovered when we applied NDC to a yeast gene expression covering 255 environmental stress and cell cycle conditions. In addition, we will also discuss the important role of genomic environment, where a particular CRE is situated, in enhancing our fundamental understanding of principles that govern context-dependent gene behaviors at the molecular level, despite the appearance of experimental data showing otherwise.

(†) Email: dnguyen@genetics.med.harvard.edu

(*) This work is supported by Alfred P. Sloan and U.S. Department of Energy Postdoctoral Fellowship in Computational Molecular Biology, and U.S. Department of Energy GTL projects.

An XML-Based File Format for Proteomic Liquid Chromatography Mass Spectrometry Data*

Dat H. Nguyen,[1‡] Kyriacos C. Leptos,[1] Leonard J. Andrews,[2] and George M. Church[1]

[1]Department of Genetics,
Harvard Medical School
77 Avenue Louis Pasteur
Boston, MA 02115

[2]Harvard Extension School
Harvard University
Cambridge, MA 02138

As high-throughput methods for studying biological systems, such as whole cell mRNA and protein studies, become standard within the scientific community, effective ways of storing data generated from these experiments become essential due to data complexity and large size. This has been partly achieved for mRNA by the MIAME standard, but not for proteomic data acquired through mass spectrometry. Like other forms of information, proteomic data need to be shared in a common, open, and transparent format in order to help both analysis and the development of computational tools. The work presented here offers our vision for a file format, ***hmsXML,*** that is size-optimal, flexible and efficient to use, simple, and based on the eXtensible Markup Language. When compared to the ***mzXML*** file format, the ***hmsXML*** file format can store the same mass spectrometry data acquired in the profile mode with 50% less space (disk storage). Open-source software suites for converting propriety liquid chromatography mass spectrometry data to the ***hmsXML*** data file, an API library for parsing it, and a GUI-based ***hmsViewer*** program for visualizing mass spectrometry data are provided freely under the GNU license agreement.

(†) Email: dnguyen@genetics.med.harvard.edu

(*) This work is supported by Alfred P. Sloan and U.S. Department of Energy Postdoctoral Fellowship in Computational Molecular Biology, and U.S. Department of Energy GTL projects.

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

Jizhong Zhou[1], Haichun Gao[1], Steven D. Brown[1], Yuri. A. Gorby[2], Mary S. Lipton[2], Heather M. Mottaz[2], Gregory E. Pinchuk[2], Xiaohu Wang[3], Yunfeng Yang[1], Soumitra Barua[1], Feng Luo[1], Jianxin Zhong[1], Xiufeng Wan[1], Liyou Wu[1], Dawn Klingeman[1], Tingfen Yan[1], Zamin Yang[1], Christopher Hemme[1], Josh N. Adkins[2], Matthew E. Monroe[2], Eric A. Hill[2], Christina L. Bilskis[2], Matthew Fields[4], Dorothea K. Thompson[1], and Timothy Palzkill[3]
Other Federation Collaborators: Alex Beliaev[2], Margaret Romine[2], Eugene Kolker[5], Kenneth Nealson[6], Joel Klappenbach[7], James M, Tiejde[7], Richard Smith[2], Carol Giometti[8], Margaret Serres[9], Monica Riley[9], Lee Ann McCue[10], and James K. Fredrickson[2]

[1]Oak Ridge National Laboratory, [2]Pacific Northwest National Laboratory, [3]Baylor College of Medicine, [4]Miami University, [5]BIATECH, [6]University of Southern California, [7]Michigan State University, [8]Argonne National Laboratory, [9]Marine Biological Laboratory, and [10]New York State Department of Health,.

Shewanella oneidensis MR-1, a facultatively anaerobic *γ-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 remain poorly understood. As a part of the *Shewanella* Federation efforts, we have integrated genomic, proteomic and computational technologies to study the regulatory networks of energy metabolism of this bacterium from a systems-level perspective.

etrA mutant characterization. In *Escherichia coli*, metabolic transitions between aerobic and anaerobic growth states occur when cells enter an oxygen-limited condition. Many of these metabolic transitions are controlled at the transcriptional level by the activities of a global regulatory protein, Fnr. In *S. oneidensis*, EtrA has been annotated as the counterpart of *E. coli* Fnr based on amino acid sequence identity. Previous characterization of an etrA insertional mutant suggested that EtrA is functional and that it participates, either directly or indirectly, in gene regulation under anaerobic conditions. In order to better understand the regulation of anaerobic energy metabolism in *S. oneidensis* MR-1, we examined the transcriptomes and proteomes of both an etrA in-frame deletion mutant and its parental strains. Both strains were individually grown in chemostats in continuous culture for 410 hours at PNNL. The growth conditions were altered from an aerobic steady state, to a microoxic steady state and to an anaerobic steady state. 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 the transition from aerobic to microoaeroxic steady states for mRNA and protein analysis. Whole genome cDNA microarrays of *S. oneidensis* that covered approximately 99% of the genome were used to elucidate the gene regulatory networks controlled by EtrA. About 20% of the ORFs showed significant differences in expression between ΔetrA and the parental strains under the steady state conditions examined, while the contents and expression patterns varied. A set of genes that were regulated by EtrA were also identified, and a conserved EtrA-binding motif was observed in their upstream regions. The sequences appear to be highly homologous to that found in *E. coli*. By studying the hybridization results for the transition samples, it appears that 150-min sampling captured transcriptional dynamics throughout the aerobic/microoxic transition. To identify members of the EtrA gene regulatory network, a set of genes whose expression was dramatically changed were further analyzed along with genes known to be modulated by EtrA. In addition, the protein expression analysis of the steady state samples was conducted using high-performance liquid chromatography mass spectrometry at PNNL. Roughly 20% of all proteins showed significant changes under the steady state conditions. While some genes are significantly changed in both mRNA and protein profiles, significant differences in expression profiles were also noticed. Overall, we believe that the combination of transcriptome and proteome analyses will allow us develop a comprehensive model for the cellular mechanism of O_2 sensing by EtrA in *S. oneidensis*.

Generation and characterization of other mutants. Shewanella Federation has taken a big effort in generating mutants for about 40 cytochrome genes in *Shewanella* by ORNL and PNNL. We are using cre-lox system to generate deletion mutants for 21 cytochrome genes. Also, about 220 transcriptional factors were identified in *Shewanella* genome based on sequence analysis. About 78 of these TFs are transcribed monocistronically or in single transcription units, and therefore, they should be good targets for insertional mutagenesis. So far, 23 insertional mutants have been generated and tagged. Preliminary work has been initiated to address metal reduction capability in these mutants, and some of them were defective in metal reduction.

Numerous other MR-1 genes have been successfully inactivated using a PCR-based, in-frame deletion mutagenesis strategy using DSP10 as the parental strain. 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, cya1-3 (adenylate cyclases), and many others. As a part of the Federation effort, we are regenerating some key regulatory mutants (arcA, crp/arcA, crp/etrA, arcA/etrA, crp/arcA/etrA, narP, narP/arcA, narP/etrA, narP/etrA/arcA) using MR-1 as the parental strain. Two Federation-level experiments to characterize these mutants have been proposed using genomic, proteomic, metabolic and computational technologies as well conventional biochemical and physiological approaches.

The regulon of the EnvZ-OmpR two-component system, which regulates porin genes in response to changes in osmolarity, was further characterized. To identify the genes that are controlled by this system, strains carrying in-frame deletion mutations in ompR, envZ, or both were constructed and investigated. Microarray analysis displayed a high degree of similarity of gene expression profiles among these mutants. Comparison of gene expression profiles of wild-type and mutant strains under various stress conditions strongly suggested that the transcription of genes SO1420 and SO1557, encoding putative outer membrane porin proteins in *S. oneidensis*, may be controlled by the OmpR-EnvZ two component regulatory system. These proteins might be the functional counterpart of the E. coli OmpC-OmpF system.

Whole proteome cloning and binding motif analysis. The first key step of developing phage display and two hybrid systems is to clone all protein-coding ORFs into universal vectors, which is a very time-consuming and tedious process. Progress as of the end of November 2004 is that 3,302 genes were cloned while no clones were obtained for 389 genes. Most of the cytochrome genes were cloned. Several key regulatory proteins such as ArcA, EtrA, Fur, and NarP have been successfully expressed in *E. coli* and more than 20 transcriptional factors were cloned into destination vectors for gene expression.

The ArcA regulon has been further analyzed. Based on microarray hybridization data and genome sequences, 254 genes were predicted to have potential ArcA binding sites. To define its binding motif experimentally, the MR-1 ArcA protein was purified from E. coli and its DNA binding activity was examined by electrophoretic mobility shift assays. To date, two ArcA interacting promoters (SO3987, a conserved hypothetical protein; SO3855, a malate oxidoreductase SfcA) have been identified from 12 candidate promoters selected according to cDNA microarray data. In addition, the minimal ArcA binding regions were determined within a 23 or 53 base pair fragment using various truncated PCR products. The ArcA binding regions contain a sequence similar (73% or 67% sequence identity) to the predicted 15-bp consensus binding motif in E. coli. The binding assay data also suggest the presence of multiple binding complexes as observed in *E. coli*. Unlike its *E. coli* homolog, however, the results indicate that phosphorylation is not required for the binding activity of MR-1 ArcA. The regulation of the malate oxidoreductase SfcA indicates that MR-1 ArcA may also play a role in regulating the TCA cycle or fermentative/energy metabolism, which has not been reported in *E. coli*. To identify binding motifs at genome level, a promoter array was designed.

Transcriptional interaction network analysis. 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. It is essential to develop computational tools to extract as much biological information as possible from ambiguous expression data with high inherent noises. Different from existing methods, we developed a computational method based on random matrix theory (RMT). 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 microarray datasets from yeast, human, *E. coli*, and *S. oneidensis*. The results demonstrated that it correctly identifies functional modules with the expected properties consistent with general network theory. Experimental validation of the predicted functions of 10 poorly characterized genes from yeast and *Shewanella* indicated that this approach can accurately predict their functions. This approach should prove useful for analyzing high-throughput genomics data for modular network identification and gene function prediction. We have applied the RMT-based methods to a variety of microarray data from multiple electron acceptor experiments, heat shock, cold shock, and etrA mutant experiments. Network maps were generated with these experimental data. The functions of about 100 unknown genes have been predicted.