

SAND REPORT

SAND2005-0790

Unlimited Release

Printed December 2004

Genomes to Life Project Quarterly Report October 2004

Grant Heffelfinger, Al Geist, Anthony Martino, Andrey Gorin, Ying Xu, Mark Daniel Rintoul, Brian Palenik

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

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

Approved for public release; further dissemination unlimited.



Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: <http://www.doe.gov/bridge>

Available to the public from

U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: <http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online>



SAND2005-0790
Unlimited Release
Printed February 2005

Genomes to Life Project Quarterly Report October 2004

Grant S. Heffelfinger
Materials and Process Sciences Center
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-0885

Abstract

This SAND report provides the technical progress through October 2004 of the Sandia-led project, "Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling," funded by the DOE Office of Science Genomes to Life Program.

Understanding, predicting, and perhaps manipulating carbon fixation in the oceans has long been a major focus of biological oceanography and has more recently been of interest to a broader audience of scientists and policy makers. It is clear that the oceanic sinks and sources of CO₂ are important terms in the global environmental response to anthropogenic atmospheric inputs of CO₂ and that oceanic microorganisms play a key role in this response. However, the relationship between this global phenomenon and the biochemical mechanisms of carbon fixation in these microorganisms is poorly understood. In this project, we will investigate the carbon sequestration behavior of *Synechococcus* Sp., an abundant marine cyanobacteria known to be important to environmental responses to carbon dioxide levels, through experimental and computational methods.

This project is a combined experimental and computational effort with emphasis on developing and applying new computational tools and methods. Our experimental effort will provide the biology and data to drive the computational efforts and include significant investment in developing new experimental methods for uncovering protein partners, characterizing protein complexes, identifying new binding domains. We will also develop and apply new data measurement and statistical methods for analyzing microarray experiments.

Computational tools will be essential to our efforts to discover and characterize the function of the molecular machines of *Synechococcus*. To this end, molecular simulation methods will be coupled with knowledge discovery from diverse biological data sets for high-throughput discovery and characterization of protein-protein complexes. In addition, we will develop a set of novel capabilities for inference of regulatory pathways in microbial genomes across multiple sources of information through the integration of computational and experimental technologies. These capabilities will be applied to *Synechococcus* regulatory pathways to characterize their interaction map and identify component proteins in these

pathways. We will also investigate methods for combining experimental and computational results with visualization and natural language tools to accelerate discovery of regulatory pathways.

The ultimate goal of this effort is develop and apply new experimental and computational methods needed to generate a new level of understanding of how the *Synechococcus* genome affects carbon fixation at the global scale. Anticipated experimental and computational methods will provide ever-increasing insight about the individual elements and steps in the carbon fixation process, however relating an organism's genome to its cellular response in the presence of varying environments will require systems biology approaches. Thus a primary goal for this effort is to integrate the genomic data generated from experiments and lower level simulations with data from the existing body of literature into a whole cell model. We plan to accomplish this by developing and applying a set of tools for capturing the carbon fixation behavior of complex of *Synechococcus* at different levels of resolution.

Finally, the explosion of data being produced by high-throughput experiments requires data analysis and models which are more computationally complex, more heterogeneous, and require coupling to ever increasing amounts of experimentally obtained data in varying formats. These challenges are unprecedented in high performance scientific computing and necessitate the development of a companion computational infrastructure to support this effort.

More information about this project, including a copy of the original proposal, can be found at www.genomes-to-life.org

Acknowledgment

We want to gratefully acknowledge the contributions of the GTL Project Team as follows:

Grant S. Heffelfinger^{1*}, Anthony Martino², Andrey Gorin³, Ying Xu^{10,3}, Mark D. Rintoul¹, Al Geist³, Matthew Ennis¹, Hashimi Al-Hashimi⁸, Nikita Arnold³, Andrei Borziak³, Bianca Brahamsha⁶, Andrea Belgrano¹², Praveen Chandramohan³, Xin Chen⁹, Pan Chongle³, Paul Crozier¹, PguongAn Dam¹⁰, George S. Davidson¹, Robert Day³, Jean Loup Faulon², Damian Gessler¹², Arlene Gonzalez², David Haaland¹, William Hart¹, Victor Havin³, Tao Jiang⁹, Howland Jones¹, David Jung³, Ramya Krishnamurthy³, Yooli Light², Shawn Martin¹, Rajesh Munavalli³, Vijaya Natarajan³, Victor Olman¹⁰, Frank Olken⁴, Brian Palenik⁶, Byung Park³, Steven Plimpton¹, Diana Roe², Nagiza Samatova³, Arie Shoshani⁴, Michael Sinclair¹, Alex Slepoy¹, Shawn Stevens⁸, Chris Stork¹, Charlie Strauss⁵, Zhengchang Su¹⁰, Edward Thomas¹, Jerilyn A. Timlin¹, Xiufeng Wan¹¹, HongWei Wu¹⁰, Dong Xu¹¹, Gong-Xin Yu³, Grover Yip⁸, Zhaoduo Zhang², Erik Zuiderweg⁸

*Author to whom correspondence should be addressed (gsheffe@sandia.gov)

1. Sandia National Laboratories, Albuquerque, NM
2. Sandia National Laboratories, Livermore, CA
3. Oak Ridge National Laboratory, Oak Ridge, TN
4. Lawrence Berkeley National Laboratory, Berkeley, CA
5. Los Alamos National Laboratory, Los Alamos, NM
6. University of California, San Diego
7. University of Illinois, Urbana/Champaign
8. University of Michigan, Ann Arbor
9. University of California, Riverside
10. University of Georgia, Athens
11. University of Missouri, Columbia
12. National Center for Genome Resources, Santa Fe, NM

Sandia and Oak Ridge National Laboratories

Genomes to Life Project

Quarterly Report

October 2004



Carbon Sequestration in *Synechococcus* Sp.:
From Molecular Machines to Hierarchical Modeling

Table of Contents

Executive Summary	7
Experimentation	9
Characterizing the Regulatory Networks and Protein Interaction Networks of a Bacterial Cell ..	10
Hyperspectral Imaging Technology	11
Expression and Purification of RuBisCO Small.....	13
Novel Data Analysis Algorithms for Tandem Mass-Spectrometry.....	14
Multi-Scale Protein Interactions Annotation Tools.....	16
Structure Determination and Analysis.....	18
Biomolecular Modeling Contribution	21
Peptide-Protein Docking Contribution.....	23
Computational Inference of Biological Networks in Cyanobacteria Genomes.....	25
Microarray Data	27
Hierarchical Simulation Platform.....	29
ChemCell: An Integrative Cell Modeling Tool.....	31
Inference of Protein Interaction Networks	32
Proteomic Toolshop	34
<i>Synechococcus</i> Encyclopedia	36
Data Entry and Browsing (DEB) Tool	38
Biopathways Graph Data Manager	40
Publications and Presentations	42

Executive Summary

This quarterly report presents the recent results and accomplishments of the Sandia-ORNL Genomics:GTL project in a format emphasizing contribution by teams of researchers in a specific topic area. In addition, the results and accomplishments since the Sandia-ORNL Genomics:GTL project was funded, the relevance to the larger project to the Genomics:GTL program, and FY05 goals are discussed briefly for each effort.

Our experimental investigations and methods development have produced significant recent results through the use of 1-D and 2-D gel electrophoresis, MALDI-TOF MS, and two-hybrid analysis to study the composition and protein interactions within the *Synechococcus* carboxysome as well as the effects of varying CO₂ levels on growth rate and protein expression patterns in *Synechococcus* Sp. WH8102. We conducted the first characterizations of the most abundant components of the proteome and characterized the phosphorus and nitrogen regulatory pathways in conjunction with computationally derived predictions of these pathways. Our hyperspectral imaging capabilities (for microarrays and whole cells) were dramatically improved as we completed a series of upgrades to significantly reduce the time for alignment/reconfiguration and reduce experimental variance by a factor of five while quantifying and removing image artifacts. We also developed an improved maximum likelihood principal components analysis (MLPCA) algorithm with speed improvements of a factor of ~5,000 for a 4,800 gene microarray while enabling analysis of 19,200 gene microarrays otherwise unable to be analyzed. Recent results of our radical new tandem MS, MS/MS data analysis methods (“Probability Profile Method,” or PPM) include prototype assignment for large and diverse data sets (~60,000 spectra) yielding a large majority of peaks identified with a surprising level of confidence. This result provides a foundation for conceptually new approaches to the analysis of MS/MS data, including de novo peptide identification with attached confidence value.

Our efforts to develop and prototype computational tools for microbial systems biology have produced recent results as well, including new computational methods for multiscale characterization of protein interactions, methods for recognizing protein functional sites, and an integrating framework for such tools. We are now working to apply these tools to other Genomics:GTL organisms in collaboration with other Genomics:GTL projects, including the ORNL-PNNL microbial proteomics effort for *Rhodospseudomonas palustris* (with F. Larimer and H. McDonald). Having previously predicted networks for phosphorus assimilation, carbon fixation, and nitrogen assimilation in WH8102, we have extended our gene regulatory network prediction efforts beyond *Synechococcus* Sp. to the phosphorus-assimilation network in *Prochlorococcus* MED, in collaboration with Penny Chisholm’s lab at MIT.

Our modeling and simulation efforts are producing structural insight into the specificity of the carbon fixing enzyme RuBisCO while producing highly relevant new simulation tools for massively parallel computing environments. The LAMMPS simulation package was officially released as an open-source parallel MD code available for download at www.cs.sandia.gov/~sjplimp/lammps.html on 9/1/2004 and has been downloaded 350 times since. We have continued to develop our ChemCell cellular response simulator which tracks spatial and temporal variations in concentrations of protein species while capturing physically relevant processes (diffusion and reaction) in cellular geometries. This quarter we enhanced the low-level algorithms used in ChemCell to perform biochemical reactions to more closely mimic the provably accurate spatial-free stochastic simulation algorithm (SSA) pioneered by Gillespie. Our hierarchical modeling work includes significant scientific and software accomplishments this quarter as well. We carried an analysis of a 13-year dataset (courtesy of W. Li, Bedford Institute

of Oceanography) of nutrient, temperature, and abundance data for picoplankton (e.g., *Synechococcus*), small nanoflagellates, and large nanoflagellates in the North Atlantic Ocean, applying a novel application of the metabolic allometric theory discussed in previous reports. This effort yielded predictions of unexpected decreases in picoplankton abundance with increasing temperature, ultimately indicating that the relative contribution of picoplankton to total carbon sequestration decreases with increasing sea-surface temperature in the North Atlantic.

We also report significant progress in our efforts to construct an integrated data infrastructure to allow advanced search and queries across a large, diverse set of data sources: sequence databases COG, INTERPRO, SWIS PROT, TIGR, JGI, PFAM, PRODOM, SMART; structure databases PDB, COILS, SOSUI, PROSPECT; pathway database KEGG; protein interaction databases BIND, DIP, MIPS; and databases of raw mass spectra and microarray data from our experimental effort and Genomics:GTL pipelines.

As discussed in previous reports, we have developed a query language and integrated schema technology to allow search and queries across diverse databases and have prototyped this capability with our “*Synechococcus* Encyclopedia,” containing all currently available database knowledge about this microbe. Our most recent accomplishments include the storage of our TIGR-fabricated *Synechococcus* full-genome microarray data in the Encyclopedia, the addition of easy-to-use web-based analysis tools for protein function characterization, protein structure prediction, comparative analysis of protein-protein interfaces, metadata entry and browsing, and electronic notebooks. Several of these tools can also transparently access supercomputers at ORNL and around the nation.

We also applied these tools to the data in the Encyclopedia to correctly predict the components of a known membrane protein – including the proteins in the membrane itself – a task that is presently impossible by experimental mass spectra analysis alone. We are now working to create an encyclopedia for *Rhodospseudomonas palustris* and *Shewanella* for use by the GTL Microbial Complexes Pipeline project and *Shewanella* Federation, respectively. We are also incorporating new pathway prediction tools developed by our Subproject 3 into the *Synechococcus* Encyclopedia, including EXCAVATOR, CUBIC, and pMAP. Finally, we completed the toolshop language, *Scigol*, designed to capture biological data-types. Efforts in this regard included the creation and debugging of the language parser, interpreter, and compiler. We also created a prototype proteomic toolshop based on *Scigol* with integrated biological function libraries written in C, Fortran, Java, and Python and graphical toolkits to allow instant inline viewing of molecular structure.

*This work was funded in part or in full by the U.S. Department of Energy's Genomes to Life program (www.doegenomestolife.org) under project, "Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org).*

Experimentation

Sandia National Laboratories

Anthony Martino, Arlene Gonzales, Yooli Light, and Zhaoduo Zhang

Our effort includes a proteomic analysis consisting of mass spectrometry-based protein identification, pull-down experiments, two-hybrid analysis, and phage display to characterize protein-protein interactions, molecular complexes, and interaction networks relevant to carbon uptake and utilization in the open-ocean cyanobacteria *Synechococcus* Sp. WH8102. To date, we have determined the protein composition and protein-protein interactions in the carboxysome, the protein complex whose main function is to fix inorganic carbon in the form of CO₂ and HCO₃⁻ to triose and other sugar and starch precursors. This work conducted within Subproject 1 provides an experimental foundation by which carbon sequestration in an organism responsible for upwards of one-half of all primary production in the world's oceans is understood. Relevance of this work as part of a bigger research community to understand the global carbon cycle and global climate change cannot be overstated. Additionally, the problem is studied in the context of developing high-throughput proteomic analysis capabilities. Finally, our research integrates computational efforts in molecular dynamic simulations to understand peptide-peptide docking (Subproject 2), protein-protein interactions and regulatory networks (Subproject 3), and whole-cell protein interaction networks (Subproject 4).

Significant recent results were presented at two conferences and have been submitted for publication (see Publications and Presentations). In summary, we used 1-D and 2-D gel electrophoresis, MALDI-TOF MS, and two-hybrid analysis to study the composition and protein interactions within the carboxysome, a protein-rich polyhedral body involved in the CO₂-concentrating mechanism (CCM), and we studied the effects of varying CO₂ levels on growth rate and protein expression patterns in *Synechococcus* Sp. WH8102. We identified the RuBisCO large and small subunits and CsoS1 and CsoS2 in a carboxysome-rich particulate fraction and thus conclude that the proteins are solely associated and highly-expressed in the carboxysome. We also determined that the carboxysome fraction contained numerous membrane-associated proteins consistent with the presence of membrane contamination. Two-hybrid analysis indicated that CsoS2 and OrfA strongly interact with a number of shell components and form dimers, and they interact with each other. To our knowledge, this is the first indication of OrfA being biochemically linked to the carboxysome. The ε-class carbonic anhydrase CsoS3 did not interact with other carboxysome components in a binary manner. CsoS3 may not interact or it may only bind in fully formed multiprotein complexes. Finally, growth rates and protein expression patterns were unchanged in the physiologically important range of 100 and 750 ppm CO₂. These experimental results currently are being compared with protein-protein interaction and network models.

In FY05, we will use high-throughput techniques to tag and synthesize all proteins known to be involved with carbon utilization. The tagged proteins will be used in exogenous pull-down experiments to more fully determine the web of protein interactions involved in carbon fixation. Additionally, we will continue using phage display techniques to understand protein interactions at the level of modular domains. Finally, we will determine CO₂ shift conditions that effect protein expression patterns. Microarray gene expression experiments will be conducted under these conditions as well to coordinate gene induction and protein expression events.

Characterizing the Regulatory Networks and Protein Interaction Networks of a Bacterial Cell

Scripps Institution of Oceanography, UCSD

Brian Palenik and Bianca Brahamsha

Our ultimate goal in this GTL project is to characterize the regulatory networks and protein interaction networks of a bacterial cell. This is particularly feasible in *Synechococcus* because of its small regulatory capacity compared to other DOE-relevant microorganisms. We will compare our results with those from other GTL projects. Determining the regulatory networks of *Synechococcus* is particularly relevant because organisms like *Synechococcus* fix much of the carbon in the oceans. Our ability to understand and predict how they will respond to global changes in oceanic conditions will depend on an understanding of how they respond to various ecological stresses.

During the last project period, we have been carrying out:

- 1) RNA isolation of mutants and wild type WH8102 under various conditions for microarray work in collaboration with TIGR and Sandia Labs;
- 2) Characterization of cells with specific gene inactivations; and
- 3) Proteome characterization of WH8102.

This work is part of Subproject 1 but interfaces with members of other Subprojects, especially work by Ying Xu at University of Georgia, Subproject 3.

Significant recent results from our project are:

- 1) The first characterization of the most abundant components of the proteome;
- 2) Identification of important “hypotheticals;”
- 3) First characterization of the phosphorus and nitrogen regulatory pathways in collaboration with Ian Paulsen, Ying Xu, and Zhengchang Su; and
- 4) Characterization and validation of computational predictions by Su and Zu.

In FY05 we will continue along the same lines, specifically focusing on a set of eleven regulatory mutants that will help define the regulatory pathways of the cell.

Hyperspectral Imaging Technology

Sandia National Laboratories

Jeri Timlin, Michael Sinclair, Howland Jones, and David Haaland

We continue to conduct experiments utilizing the hyperspectral imaging technology in support of the Subproject 1 goals to develop a fundamental understanding of *Synechococcus* regulatory networks using state-of-the-art measurement science. Our efforts in FY04 have been focused on:

- 1) Upgrading the hyperspectral imaging system for increased flexibility and ease of operation;
- 2) Improving the quality of the hyperspectral data and multivariate data analysis results;
- 3) Providing quality control for existing TIGR microarray manufacturing; and
- 4) Identifying target areas where our technology may impact other GTL projects (current and future).

Our work is part of Subproject 1 but the microarray efforts are heavily integrated with Subproject 3. Our contribution is significant to Sandia's GTL project because the understanding of the biological relationships within this project is dependent on the quality of the data generated. We are striving to ensure that the most accurate microarray data is produced and our instrument is ready for the increased challenges of future GTL projects. The impact of hyperspectral scanning should be realized by a broader community as it can potentially improve many fluorescence imaging applications by providing independent maps of all emitting species in a sample leading to improved reliability and accuracy of the data, improved throughput, and increased dynamic range.

Over the course of this project we have used our hyperspectral scanning technology to:

- 1) Identify problems with our own TIGR *Synechococcus* microarrays and *Shewanella* microarrays from other GTL efforts;
- 2) Demonstrate the potential for increased throughput using arrays labeled with four dyes;
- 3) Expand our 2-D imaging capabilities to include easy choices of excitation wavelength and spatial resolution; and
- 4) Improve our measurement reliability and accuracy with positioner, spectrometer, and detector upgrades.

Our work with the GTL microarray data has led us to more fully understand our multivariate analysis algorithms and improve them.

Our most recent accomplishments have improved our instrumentation and software analysis tools. We have completed a series of upgrades to the sample X-Y stages and optomechanical hardware to allow us to significantly reduce the time for routine alignment and system reconfiguration and reduce experimental variance by a factor of five. We identified our spectrometer as the limiting component in the system and have designed and fabricated an optimized spectrometer with minimal image aberrations. We have developed software tools to assist in quantifying the presence of image artifacts and their detrimental impact on the hyperspectral data and to remove these artifacts from our data. In addition, we have begun interactions with Wim Vermaas at ASU to image and discern native pigments in *Synechosystis*. Our team is also organizing and uploading the TIGR *Synechococcus* full-genome microarray data to the ORNL-GTL server and providing archive storage of the actual microarray slides.

In FY05 we will continue to support the SNL GTL microarray efforts by improving our hyperspectral imaging system capabilities and our powerful multivariate analysis tools. We will completely integrate our improved spectrometer and new detector, which will provide higher signal throughput, improved image acquisition speed, and greater data quality. We will diligently pursue opportunities for additional GTL applications (both within and external to GTL) that could benefit from hyperspectral imaging technology.

Expression and Purification of RuBisCO Small

University of Michigan

Hashim Al-Hashimi, Shawn Stevens, Grover Yip, and Erik Zuiderweg

Our efforts in FY04 have been focused on expression and purification of the small subunit of RuBisCO (rbcS), the chosen target for high-throughput screening, and optimization of relaxation pulse sequences to characterize protein dynamics by NMR spectroscopy. This work is included in Subproject 1 and is associated with the experimental work ongoing in Tony Martino's laboratory at Sandia as well as computational methodology developed in collaboration with Andrey Gorin of ORNL and in the future, Diana Roe of Sandia, both of Subproject 2. Our goal is to develop and apply NMR methods for high-throughput characterization of protein-protein interactions. This work is to be integrated with other methodologies, such as phage display and computational modeling, to obtain information with a speed, robustness, and resolution that could not be attained by any of these techniques alone.

Since the inception of the project, and in close collaboration with the group of Andrey Gorin at ORNL, we have refined the dipolar couplings-based combinatorial assignment program (CAP) for high-throughput backbone resonance assignments of proteins; the rate-limiting step in NMR characterization of protein-protein interactions. This novel high-throughput assignment strategy has been successfully tested on model proteins. We have also developed improved NMR pulse sequences to measure conformational dynamics in protein-protein binding sites. The new NMR pulse sequences developed primarily in the Zuiderweg group have increased the temperature stability, reliability, and reproducibility of the experiment, which is critical for our goals of identifying protein binding sites on the basis of pre-existent induced-fit dynamic equilibria in a high-throughput manner. A paper has been published on these results (see Publications and Presentations).

We are now applying these approaches on rbcS and other proteins relevant to the carboxysome. NMR spectroscopy is arguably the most amenable tool in structural biology for resolving biomolecular complexes to high resolution. Therefore, optimizing methodology to obtain high-resolution information on protein-protein complexes rapidly is valuable far beyond the scope of this project.

We have successfully expressed and purified both unlabeled and ^{15}N -labeled rbcS and are currently in the process of optimizing the yield. Our preliminary spectra for rbcS are very promising, displaying the desirable chemical shift dispersion characteristic of a folded well-behaved protein. In the first quarter of FY05, we plan to finish sample conditioning of rbcS in preparation for obtaining backbone assignments. This will include optimization of buffer, pH, temperature, and protein concentration, as well as incorporating minor changes to the rbcS construct. We will then test our new assignment strategy (CAP) against the traditional procedures to validate the process on a previously unassigned protein. By the end of the second quarter, we anticipate incorporating peptides identified by Tony Martino's laboratory into the study. Restraints we provide will be sent to Diana Roe for modeling. In parallel, we will also begin to pursue the putative PdZ domain (SYNW1175), which has been identified by Diana Roe as an attractive target for computational modeling of protein recognition in Subproject 2.

Novel Data Analysis Algorithms for Tandem Mass-Spectrometry Oak Ridge National Laboratories

Robert Day, Nikita Arnold, Andrei Borziak, Nagiza Samatova, and Andrey Gorin

This effort is part of the Subproject 2 (“Molecular Machines”). Tandem mass spectrometry (MS/MS) is the leading GTL high-throughput methodology for identification and characterization of the protein mixtures. The existing MS/MS data analysis algorithms have fundamental problems (confidence of protein identifications could be derived only empirically, there are no feasible algorithms for a number of complex identifications, such as cross-linked species). All experimental GTL projects are heavily deploying mass spectrometry for their most important experimental pipelines: verification of the protein samples, capturing whole-cell proteome images, and determination of protein molecular machines. The development of MS/MS data analysis algorithms with new capabilities is considered a crucially important mission for this Genomics:GTL Goal 4 project.

We have proposed a radically new approach to tandem MS MS/MS data analysis. The principal idea can be described as a probabilistic labeling of the individual peaks or as a detailed analysis of the spectra leading to peak separation into specific categories (b-ion, y-ion, double-charged b-ion, etc.). In our experiments, Probability Profile Method (PPM) assignment was conducted on large and diverse data sets (~60,000 spectra). The large majority of peaks were identified by PPM with a surprising level of confidence, providing the foundation for a whole set of conceptually new approaches to the analysis of MS/MS data, including de novo peptide identification with the attached confidence value.

Toward the realization of these ambitious goals, the following milestones were achieved during FY03 and FY04:

- 1) Development of the libraries/toolsets for reading MS/MS spectra, spectra manipulation, ion generation, statistics collection, and other utilities;
- 2) Formulation of Bayesian approach to peak identification, which integrated a multitude of weak signals useful for peak categorization and identification, and development of the corresponding libraries and software tools;
- 3) Construction of de novo peptide labels and collection of the relevant statistics;
- 4) Implementation of a novel charge determination algorithm and its successful use in the ORNL-PNNL GTL pilot facility; and
- 5) Development and testing of the PPM-Chain program, which combines de novo labels and database search. The PPM-Chain performance could be positively compared to Sequest, the current “gold standard,” but it also adds a number of additional capabilities, which are opening new horizons for deep analysis of the rich proteomic data content. The section (e) was mostly completed during last reporting period.

We are directly working with several GTL Centers: George Church and his postdocs at HMS, Gordon Anderson at PNNL, and Hayes McDonald at ORNL. We also received very exciting offers to collaborate from the University of Georgia, Harvard University, and the University of California at San Diego.

In FY05 we plan to open two high-performance servers. The PPM server will provide probabilistic peak identifications for the spectra supplied by the community and will allow further research into potential new applications of PPM-derived probabilities. PPM-Chain will be the on-line provider of de novo labels for all difficult cases of sample dissection and should become

another important GTL community resource. We also plan research into mathematically rigorous confidence values for the peptide/protein identifications.

Multi-Scale Protein Interactions Annotation Tools

Oak Ridge National Laboratories

Praveen Chandramohan, Pan Chongle, Al Geist, Andrey Gorin, Ramya Krishnamurthy, Rajesh Munavalli, Byung Park, Nagiza Samatova, and Gong-Xin Yu

This effort aims to develop computational tools for a systems-level annotation of protein interactions. Our goal is to provide a fresh stream of computational technologies capable of discovering strong functional clues from complex and noisy data, and then integrating those clues to enable systems-level annotation of biomolecular interactions. As a result, this project will deliver a capability that would allow researchers to determine a protein function and put it in both a macro context of interaction networks and a micro context of particular residues involved into functions.

To date our major accomplishments include:

1. *Computational tools for multiscale characterization of protein interactions.* Given protein primary sequence information, our computational tools predict with a reasonably high confidence: (a) whether two proteins are interacting; (b) if two proteins are interacting, then what their interface residues (docking interfaces) are; and (c) what the contacting residue pairs are. These predictions are being integrated into a framework for reconstruction of protein interaction networks.
2. *Computational tools for recognizing protein functional “spots.”* Given a multiple sequence alignment of a protein family, our computational tools identify clusters of residues whose structural, dynamic, and physicochemical properties directly or indirectly affect protein interactions, and hence, a protein function. Specifically, such residue clusters include residues vital for enzyme-substrate, domain-domain, and regulator-enzyme interactions.
3. *A framework for high-resolution functional annotation of protein machines.* Our tools could improve current annotations of microbial genomes as follows: (a) providing annotations at higher specificity level, that is, not only identifying the protein family but also identifying a specific substrate or even substrate-specificity determining residues; (b) providing functional annotation for hypothetical or conserved hypothetical proteins; (c) identifying the interacting proteins vital for a protein function; and (d) identifying critical genes for a given biomolecular pathway.

In addition to the specific accomplishments outlined above, our research presents state-of-the-art advancement in the field of computational identification and functional annotation of protein interactions. We provide a unique capability that any other institution can hardly offer. Thus, this project not only addresses critical computational needs of the DOE Genomics:GTL but also provides benefits to a large biological community by making our tools available for download or through a web interface.

Characterization of protein machines (the focus of this project) is one of the key goals of the DOE Genomics:GTL Initiative and the Subproject 2 of the Sandia-ORNL Genomics:GTL project.

Within the GTL Centers, our tools are or will be applied as follows:

1. *Characterizing protein interactions in Synechococcus Sp. (SGTL) and R. palustris (F. Larimer and H. McDonald).* By using MS pull-down experimental data from M. Buchanan’s

Center, we have demonstrated (on several target protein complexes) the potential of our tools to make predictions beyond the experimental capabilities, for example, membrane protein interactions. Likewise, we showed that the tools could indeed be used for identifying the targets for MS experiments. The work is underway to apply our tools to *R. palustris* and *Shewanella* on a genome-scale.

2. *Refining and facilitating the functional annotations of microbial genomes.* This re-annotation effort is currently being done in close collaboration with Brian Palenik, Ying Xu, and Charlie Strauss (on *Synechococcus* Sp.), as well as with Frank Larimer and Hayes McDonald (on *R. palustris*). Brian and Frank bring connections to the rest of the annotation community.

3. *Facilitating biophysical characterization of target protein complexes, such as RuBisCo.* We have applied our tools to identify residue clusters responsible for the CO₂/O₂ specificity of the RuBisCo enzyme, a target enzyme of the Sandia-ORNL Genomics:GTL project. These predictions can potentially reduce the search space for biophysical characterization conducted by Paul Crozier, Sandia, Subproject 2.

4. *Improving the performance of structure prediction and molecular docking tools, such as ROBETTA and PDOCK.* This work will be conducted in close collaboration with Charlie Strauss, LANL and possibly Diana Roe, Sandia, both of Subproject 2.

Our major deliverables will include curated annotation of protein interactions in *Synechococcus* Sp. and *R. palustris* and refined functional annotations of these target genomes. This effort will be conducted jointly with our collaborators listed above. We will refine and package our prototyped tools to make them available for a large community of biologists. A list of more detailed deliverables towards achieving these major goals is available upon request.

Structure Determination and Analysis

Los Alamos National Laboratory

Charlie Strauss

There are five threads of my current work that will become the dominant part of my FY05 efforts. Many of these are collaborations with other members of this Genomics:GTL project, some of which I lead and others that are shared efforts. These are: (1) continued development, maintenance, and application of the ROBETTA Protein Structure prediction server; (2) structure based annotation; (3) developing methods suited to directed protein structure sampling; (4) data fusion; and (5) contributions to Andrey Gorin's de novo mass spectroscopy analysis platform.

1) The ROBETTA structure prediction server is a fully automated web-based interface to the Rosetta Structure prediction system. The user merely provides a sequence and the 140-node cluster parses the sequence into domains and provides either homology models or ab initio structural models, as appropriate. The latter functionality is its unique feature. Current work on system improvements includes development efforts to provide a ported Live CD version of this that users could install on their own systems. The Live CD system work is just starting and there are both technical and legal hurdles.

2) The ROBETTA pipeline includes a simple structure-based annotation. Predicted structures are compared to known structures and ranked by the probability of that similarity occurring by chance. We have been exploring ways to convert these rankings to confidence statistics. This can be done either by considering other structure-quality factors besides structural overlap or by utilizing external information. In the former case, we are working with the Baker Group at the University of Washington. In the latter case, we are developing a fast literature screening method to compare two bodies of literature in an automated fashion. This will allow us to quickly sieve the literature for confirmation of high-confidence matches found in the first step.

3) Modeling of protein structure hangs critically on controlled exploration conformational space. To give three examples: a) the conformational space is ludicrously larger than any theoretical computer could exhaustively explore; b) when proteins dock or bind ligands, small changes in both rotamers and backbones occur; and c) in the reverse problem of given a backbone, discover what alternative sequences are compatible with it. To reach higher resolution, we need to improve both the global conformational search to direct it to under sampled regions of topology, as well as to more finely sample the local topology. This work requires investigation of methods suited to massively parallel computational scaling.

In this regard several threads are being explored. One initiated by Andrey Gorin and performed mainly by Yaohang Li involves adapting Rosetta to a parallel tempering algorithm. For refined local structure, we are also looking at exhaustive enumeration of perturbative searches (led by Martin Tomba at the University of Washington). Finally, Chenfeng Huang and myself are working with David Wolpert at NASA Ames to develop a combinatorial optimization method suited to sequence remodeling.

4) In addition, we are also looking for ways of fusing other sorts of information developed in the Sandia-ORNL Genomics:GTL project into improving both the accuracy of models and, perhaps more importantly, our confidence in model predictions and structure-based annotations. We plan to incorporate sequence-based contact predictions. Two other investigators in this project are

capable of generating these: Nagiza Samatova (based on her correlated mutation analysis) and Jean -Loup Faulon (beta strand pairing based on his peptide docking work).

5) We recently have developed a Hidden Markov model that allows us to discover the most probable set of ions in a mass spectrum that can be linked by a feasible peptide. The method works well and we are refining it to practice in collaboration with Andrey Gorin's mass spectrometry project.

Expected near-term products are papers on:

- 1) Parallel-tempering analysis for Rosetta;
- 2) Hidden Markov analysis of mass spectral data, and publication of computer algorithms;
- 3) Literature-screening recognition of functional families; and
- 4) Transfer of ROBETTA server annotation to ORNL based project wide annotation system.

Medium term:

- 1) Improved confidence intervals in protein structure-based annotation;
- 2) Incorporation of this into the ROBETTA server to provide true confidence intervals;
- 3) Incorporation of this into the Mammoth algorithm;
- 4) Direct transfer mechanism for dumping SQL tables from ROBETTA to ORNL *Synechococcus* server (led by Nagiza Samatova); and
- 5) Live CD ROBETTA or a streamlined port.

Longer term:

- 1) Contact constraints on structure prediction: with Jean-Loup Faulon or Nagiza Samatova;
- 2) Use of these as screening for ROBETTA server predictions; and
- 3) Incorporation of literature screening into ROBETTA.

Many of the topics we are working on not only directly integrate activities in this Genomics:GTL project but they also position us for upcoming Genomics:GTL facility work. In particular, sequence redesign impacts Facilities 1 and 2, since protein production and characterization require protein sequence redesign for improved solubility, stability, and expression. While we cannot solve all those in this year's effort, our explorations of combinatorial sequence redesign is an enabling component. Our approach is also intended to scale in a massively parallel fashion, thus making it attractive to traditional OASCR project interests as well. The data fusion aspects of our work that include using sequence-based contact predictions for structure analysis fits in well with Facility/Goal 4 Genomics:GTL project needs. In the future, proteomics pipelines will be producing large quantities of data capable of enhancing structure predictions and for which structure models can be used to explain various functional inferences from both experimental and computational high-throughput techniques. The work on this project seeds this ground by developing methods to explore more varied protein topologies, as well as to use limited constraints to either confirm or improve structure prediction.

We are actively transferring our developments to an eager ROBETTA user base, which spans dozens of countries and hundreds of academic and government labs. Improvements to our structure comparison, literature analysis, and structural modeling have an assured pathway for use. The demand for our services can only increase: gene discovery is 65-times faster than experimental structure determination. Currently, it would take 300 years to determine the structure of presently known sequences by X-ray/NMR methods. Computational methods for structure analysis thus will be in growing demand. Most of the structure analysis community is focused on the homology modeling problem rather than the ab initio modeling problem of protein

without recognizable homology to known structures. As more X-ray structures become available, the difficulty of homology modeling fades, but the importance of modeling the variable regions of these, notably variable loops and conformational changes in binding, will grow. The latter are ab initio problems. The limiting feature will thus become the development of scalable computational methods of exploring conformational space.

Biomolecular Modeling Contribution

Sandia National Laboratories

Paul Crozier and Steve Plimpton

The goal of this work is to understand the structure and function of the RuBisCO enzyme, which plays a key role in carbon fixation in *Synechococcus* and a variety of other photosynthetic organisms. Specifically, we hope to correlate the carboxylation rate and selectivity of the RuBisCO enzyme to its structure and amino-acid composition. More broadly, we are developing atomistic modeling tools and algorithms that are of general utility for biomolecular modeling of proteins and protein complexes. This work is connected to the experimental work of Subproject 1 that is characterizing the composition of the carboxysome, which includes RuBisCO and its possible binding partners. It is also complementary to the Subproject 4 work to model the carbon fixation pathway in *Synechococcus* at a cellular level, since the output of that pathway is sensitive to the biochemistry of RuBisCO.

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

We also worked this quarter to add several algorithms to our molecular dynamics code LAMMPS, which we use to model RuBisCO. The algorithms were of two types: those that speed-up the modeling and those that increase its sampling efficiency for rare events such as gate openings and closings. Specifically, we added a general rRESPA capability to do hierarchical timestepping, a precomputation of Coulombic forces stored in tables, options to freeze portions of the model as rigid bodies and avoid unnecessary force computation, and the TMD algorithm mentioned above that enables portions of the gate to be rapidly opened and closed. The LAMMPS simulation package was officially released as an open-source parallel MD code available for download at (www.cs.sandia.gov/~sjplimp/lammps.html) on the first of September. Since then, it has been downloaded 350 times.

In FY05 we will explore a variety of RuBisCO structures with models of varying levels of sophistication and cost, and compare the TMD-predicted enzyme performance to the experimentally measured performances. We will also continue to develop algorithms in support of more efficient and effective modeling and incorporate them into the publicly available version of LAMMPS.

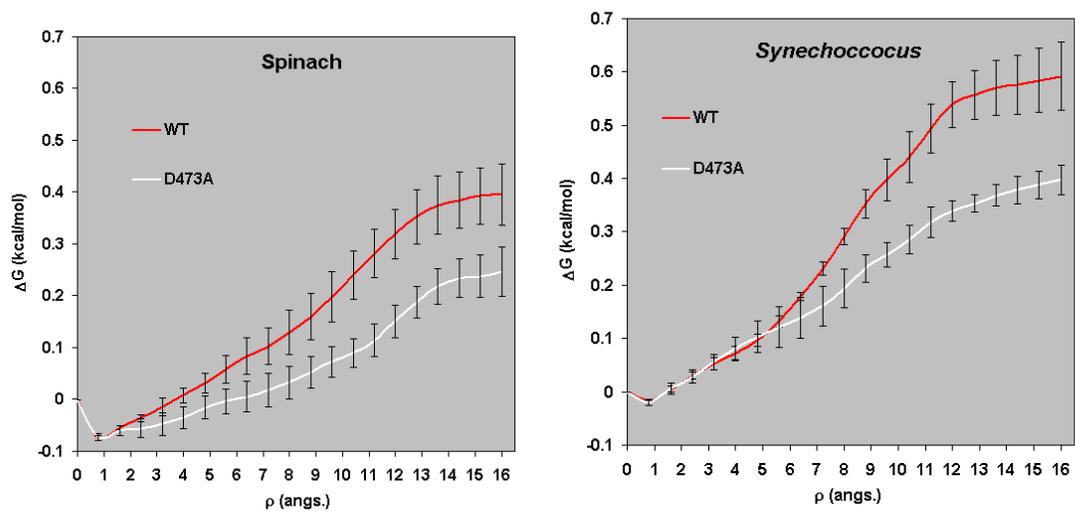


Figure 1. Change in free energy of RuBisCO gate opening from TMD simulations. As the distance from the closed structure (ρ) is increased, the free-energy profile is computed.

Peptide-Protein Docking Contribution

Sandia National Laboratories

Diana Roe and Bill Hart

In our efforts to create a toolkit of docking algorithms and capabilities that can be applied to a variety of peptide-protein and protein-protein interactions, we have completed a prototype of such a tool, PeptideDock, designed to aid experimentalists in studying protein-protein interactions through protein-peptide interactions.

Many protein-protein interactions, especially in intracellular signaling complexes, are mediated by modular peptide-binding domains, for example, PDZ, SH3, SH2, WW, PTB, FHA, that may interact with different protein-binding partners. The complementary interaction domain on the binding partners is often a linear stretch of protein, such as a carboxylate tail in the case of SH3 domains, as seen in Figure 2. The binding sequence of these partners usually consists of a common core motif with binding specificity in the amino acids flanking the core sequence. Biologists use experimental techniques like phage display to identify peptide-binding domains, from which they can identify a consensus sequence pattern for the core peptide sequence and a specific set of sequences matching the consensus pattern, demonstrating the binding specificity for that protein.

PeptideDock starts with a backbone conformation of a peptide binding in a peptide-binding groove, and evaluates **all** possible peptide sequences for that backbone. The backbone is allowed to move locally, and all the sidechains are fully flexible using a combination of discrete rotamer libraries for full conformational sampling, followed by local minimization of each rotamer conformation. This quarter we expanded the evaluation function and developed a combinatorial enumeration technique that is used to enumerate all low-energy peptide sequences. Specifically, we:

- 1) Added a penalty for placing sidechains onto a particular backbone phi/psi conformation, based on the probability of finding that sidechain on that backbone phi/psi within the backbone-dependent rotamer library;
- 2) Developed a Boltzmann model to describe the interaction energy of each residue type at each position, based on the Boltzmann-weighted distribution of all the rotamers and using a mean-field theory approximation;
- 3) Extended the PICO optimization library to perform combinatorial enumeration, for example, to enumerate all solutions within a tolerance of the best solution; and
- 4) Developed combinatorial lower-bounding techniques and a customized branch-and-bound search.

We have begun testing this solver on a variety of datasets taken from the PDB, for which these enumerations can often be performed within seconds and at most within a few minutes.

We have discovered that the core technology developed for PeptideDock can be extended to several related modeling tools that can be used to answer important biological questions about how proteins interact with each other within a cell. In FY05 we will continue supporting this Genomics:GTL effort in developing tools to understand protein-protein interactions by completing our PeptideDock tool and beginning work on these new tools.

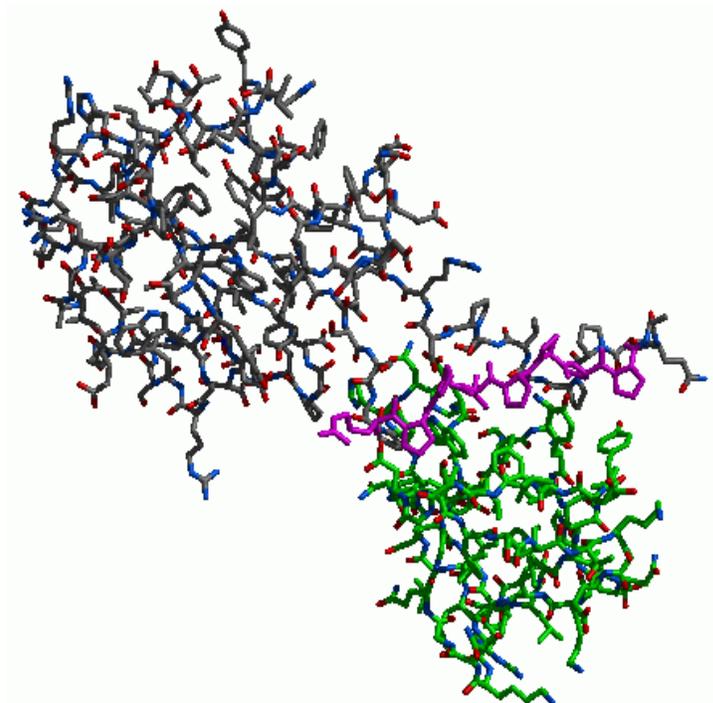


Figure 2. Complex of the SH3 domain from tyrosine protein kinase (green) complexed with P38 (grey), with linear portion highlighted (magenta).

Computational Inference of Biological Networks in Cyanobacteria Genomes

University of Georgia¹, University of Missouri at Columbia², University of California at Riverside³, University of California at San Diego⁴

Zhengchang Su¹, PguongAn Dam¹, HongWei Wu¹, Victor Olman¹, Xiufeng Wan², Xin Chen³, Dong Xu², Brian Palenik⁴, Tao Jiang³, and Ying Xu¹

In the past quarter (July to September 2004), we continued our work in the prediction of detailed models of a few response networks in *Synechococcus* Sp WH8102 and *Prochlorococcus* MED, and in the development of computational tools to facilitate such prediction work. Our work represents the core part of the Subproject 3, which addresses the key issues in computational elucidation of biological networks.

Key ideas of our approach to elucidation of biological networks are to:

- 1) derive genes involved in the core components of a target network (metabolic, signaling, regulatory, or a whole response network) through functional predictions of genes and predictions of operons (and microarray data when they are available); and
- 2) predict physical and functional interactions among these components through prediction of protein-protein interactions, protein-DNA interactions, and transcription regulatory elements and signaling systems, for example, sensor-regulator two-component systems, leading to the initial prediction of the core network structure.

Further refinement and expansion include recruiting additional elements through prediction by the rule of “guilt-by-association.” One of the key sources of information, when lacking microarray gene expression data, is all the genomes that have been sequenced and publicly available. Our experience has demonstrated that a great amount of structural and functional information about biological networks could be derived through careful comparative genome analysis. Using such information, in conjunction with limited experimental data mainly extracted from the literature, we have predicted sound models for the carbon fixation response network and phosphorus assimilation network for *Synechococcus* Sp. WH8102. These predicted models are biologically sound and consistent with our current knowledge about these biological processes. They also provide valuable information that fills the gaps in our knowledge about these biological processes, which could be used to guide further experimental investigation. Brian Palenik’s lab has conducted a number of experiments based on the computational predictions. We expect that these experiments will validate our predicted models. Recent experimental data (mass spectrometry) collected by Brian’s lab has clearly helped to speed up our recent prediction work on the nitrogen-assimilation network, which is being finished.

Over the course of the project, the most significant results of our research could be summarized as follows:

- 1) We have developed a highly effective computational protocol for prediction of biological pathways and networks, which is particularly effective when there is limited experimental data;
- 2) We have predicted models of three important response networks in *Synechococcus* Sp. WH8102, which are the phosphorus-assimilation network, the carbon-fixation networks; and the nitrogen-assimilation network, through the application of this computational protocol;

- 3) We have extended our prediction capability to other organisms, including the phosphorus assimilation network in *Prochlorococcus* MED through a collaboration with Penny Chisholm's lab at MIT;
- 4) We have developed a strong capability for inference and prediction of functional modules in microbes in general, which provides a very important piece of information for biological network inference (two papers, which are not listed in the Appendix, are currently under preparation);
- 5) We have developed a suite of computational prediction tools which facilitate the network inference work, which include (a) a highly effective tool for prediction of operons in microbe, (b) a very effective prediction tool for identification of cis regulatory elements called CUBIC, (c) a prediction program, EXCAVATOR, for analysis of microarray gene expression data, (d) a prediction program for de novo sequencing method for identification of proteins, (e) a prediction program for protein-protein interactions, and (f) a prediction program for identification of terminator signals in an operon (the first four programs have been used by a number of other research labs); and
- 6) We have developed a database for predicted data for *Synechococcus* Sp. WH8102 at UGA.

All this prediction and tool development work has led to over twenty research papers that are either published, in press, or submitted for publication (see Publications and Presentations for the full list).

For FY05, while we will continue to make more network predictions, refine our computational prediction protocol, and improve and further develop more prediction tools, we will also add new dimension to our research work through the study and characterization of the dynamic behaviors of our predicted network models, mainly to identify chokepoints and critical pathways, which could be used to guide genetic engineering to improve the performance of certain biological processes in microbes that could be used for carbon fixation or bioremediation.

Microarray Data

Sandia National Laboratories

David Haaland, Ed Thomas, and Chris Stork

We continue to focus on robust statistical analysis of microarray data, improving microarray data quality, obtaining models of experimental microarray data, generating realistically simulated microarray data, and developing new multivariate methods for improving the analysis of microarrays for which correlated and/or heteroscedastic errors are present. This work supports Subproject 3 goals to develop a fundamental understanding of *Synechococcus* regulatory networks and to verify computer models of regulatory networks. Efforts in FY04 have focused on:

- 1) Designing optimal print designs for the TIGR whole-genome *Synechococcus* microarrays;
- 2) Evaluating the quality of resulting TIGR microarray data;
- 3) Using replicate microarray data to understand and model the error structure of TIGR *Synechococcus* microarrays;
- 4) Generating realistic simulations of microarray data using representative signal and experimentally derived models of error covariance;
- 5) Developing more efficient maximum-likelihood analysis methods that incorporate the microarray error structure in analyses; and
- 6) Providing robust statistical analysis of the data and feedback to TIGR collaborators on best statistical practices.

This contribution to the Sandia-ORNL Genomics:GTL project improves the quality and reliability of the analysis of TIGR microarray data and increases the sensitivity to differentially expressed genes while simultaneously decreasing false positive rates. Over the course of the project, we have demonstrated that properly designed microarray experiments can identify experimental parameters that limit data reproducibility and reliability, and we have demonstrated a factor of 3 \log_2 unit improvement in repeatability of printed yeast microarrays. We have shown that realistically simulated microarray data can be productively used for evaluation of competing feature selection and multivariate microarray analysis algorithms. Although spatially correlated errors currently limit the reproducibility of most microarray data, our new efficient maximum-likelihood principal component algorithms can reduce the impact of these spatial errors on microarray analysis. We have performed robust statistical analysis of the whole-genome microarray data to feed into the computational component of Subproject 3.

Our most recent accomplishments include the generation of realistically simulated whole-genome *Synechococcus* microarray data from experimental replicate microarray data and a demonstration that our improved maximum likelihood principal components analysis (MLPCA) algorithm can be applied to realistically simulated data. Speed improvements over conventional MLPCA are a factor of ~5,000 for a 4,800 gene microarray, and the algorithm allows analysis of 19,200 gene microarrays that cannot be analyzed with conventional MLPCA algorithms. We have also provided robust analysis of TIGR microarray data that compare gene expression of *Synechococcus* under standard conditions and co-growth with vibrio bacteria, low PO_4 levels, and NH_4 as the nitrogen source.

In FY05, we will apply the MLPCA methods to experimental data, continue support for the robust statistical analysis of TIGR microarray data, and identify methods to improve TIGR

microarray repeatability. We will extend our data analysis to high-fidelity whole-cell imaging. We will analyze our newly constructed 3-D confocal hyperspectral fluorescence images of live wild-type and mutant *Synechocystis* to obtain quantitative representations of the concentrations of each of the native pigments in these photosynthetic bacteria. These latter studies will position us for the FY'06-08 GTL proposal. This research and the new algorithms developed for microarray analysis and elucidation will benefit all GTL projects with microarray analyses. In addition, the demonstration of 3-D hyperspectral imaging at the diffraction limit for *Synechocystis* will show the power of this new method to quantitatively elucidate the highly spectrally and spatially overlapped pigments present in bacteria. These methods can be extended to fluorescently labeled prokaryotic and eukaryotic cells and to kinetic investigations of fluorophores to determine both reaction and mechanisms reaction rates.

Hierarchical Simulation Platform

National Center for Genome Resources

Damian Gessler and Andrea Belgrano

The hierarchical modeling work of the project accomplished significant scientific and software results since the last progress report. Scientifically, we analyzed a 13-year dataset (courtesy of W. Li, Bedford Institute of Oceanography) of nutrient, temperature, and abundance data for picoplankton (*e.g.*, *Synechococcus*), small nanoflagellates, and large nanoflagellates in the North Atlantic Ocean as part of our modeling work. We applied a novel application of the metabolic allometric theory reported in our earlier progress reports and discovered significant and antagonistic responses of picoplankton and large nanoplankton to shifts in their respective temperature-corrected resource flux estimates. The analysis showed that as temperature increases, there is a shift in the species composition of the community such that picoplankton abundance decreases. This was unexpected, as the classical expectation is that picoplankton numbers would increase with temperature. To explain this, we are performing modeling work (see below) to examine the nonlinear response to temperature experienced by all community members that affects growth/mass trade-offs nonsymmetrically. The data show that total community energy-flux remains invariant with respect to cell mass, in agreement with predictions of the energy equivalent rule. The result – and major relevance for Genomics:GTL – is a decrease in the relative role that picoplankton contribute to total carbon sequestration upon increases in sea-surface temperature in the North Atlantic.

The modeling platform simulates years of *Synechococcus* growth via modeling carbon sequestration over order-of-magnitude differences in time scales (subcellular to ecosystem time-scale processes) and through numerous hierarchical levels. We use the abiotic forcing of changes in atmospheric irradiance to drive changes in ocean nutrient concentrations, cellular carbon fixation, carboxysome photosynthesis, population growth, and back to changes in ocean nutrient concentrations. We accomplish this by using a novel event-driven, disparate-time modeling platform of our own development funded under this Genomics:GTL project. The modeling platform handles multiscale, disparate-time processes, such as cellular carbon fixation embedded in growing populations under seasonal changes in irradiance, by passing control to only those nodes that require computation as parameter values change or as scheduled events are fired/delivered. This creates a type of application-driven time slicing that distributes load dynamically as the state of the simulation changes. The platform coordinates parameter dependencies between levels, such as how cellular nutrient quota uptake rates are dependent on ocean nutrient concentrations, temperature, and irradiance, and automatically brokers events between broadcasting and listening nodes and assures parameter and event temporal integrity. Using this approach we simulate the embedding of diurnal carbon fixation “pumping” (caused by the oscillating availability of light every 24 hours) within a much slower seasonal oscillation in irradiance caused by the earth’s rotation. An important property of this approach is that the modeling allows data-centric processes to be embedded along with analytical equations. One direction we are currently exploring is how to use easily and inexpensively available data in some parts of the simulation to drive and guide modeling predictions in more interesting parts of the simulation. We are exploring a simulation approach to complex systems that uses data to dynamically and iteratively check and set bounds to the more speculative parts of the simulation and exploits the use of surrogate data at one hierarchical level to increase the predictive power of other hierarchical levels.

This is relevant to the Sandia-ORNL Genomics:GTL project because it links *Synechococcus* cellular carbon fixation to the oceanic carbon budget and vice-versa. This is the only work in the Sandia-ORNL Genomics:GTL project that addresses the role of *Synechococcus* carbon fixation in the larger picture of ecosystem consequences. In addition, this work is yielding a new modeling and simulation approach specifically designed for hierarchical, disparate-time simulation. Over the course of this project, we have designed, developed, implemented (coded), tested, and applied a hierarchical disparate-time modeling platform that incorporates extended stoichiometric and allometric theory. We have applied this capability to new applications in carbon sequestration and discovered new insights into *Synechococcus* carbon sequestration as a function of abiotic forcing and abundance/mass trade-offs. The biologically relevant results from this effort include evidence of a decreasing role of picoplankton in total sequestered carbon in the North Atlantic as temperature increases. In addition, we have produced modeling results of embedded seasonal and diurnal oscillations in *Synechococcus* carbon sequestration as a function of irradiance, as well as an increasingly mature hierarchical, disparate-time modeling and simulation platform.

The impact of this effort on the larger Genomics:GTL program is to extend our understanding of the dependence of marine carbon sequestration on changes in sea-surface temperature while delivering a new multiscaling simulation platform and new data-driven simulation techniques. More broadly, this effort is yielding fundamental knowledge of how changes in irradiance and temperature affect marine species composition, to the possible detriment of picoplankton (*Synechococcus*) population sizes

In FY05 we will publish our scientific results and the modeling platform used to obtain them. Our model development focus will be to mature the model scientifically by the explicit inclusion of flagellates in a multispecies model and to explore the application for general complex system modeling, using data in one part of the simulation to drive and check predictions in other parts.

ChemCell: An Integrative Cell Modeling Tool

Sandia National Laboratories

Steve Plimpton and Alex Slepoy

In recent months, we have continued to develop our ChemCell simulator, which tracks the spatial and temporal variations in concentrations of protein species as the molecules diffuse within a cellular geometry and biochemically react with each other. Last quarter we reported on the development of an initial ChemCell model for the network of reactions in *Synechococcus* that converts CO₂ to organic carbon compounds (glucose), mediated by the carboxysome RuBisCO enzyme. This effort is a collaboration with the Subproject 1 experimental characterization of *Synechococcus* carboxysomes as well as with the Subproject 2 effort to perform atomic-scale modeling of RuBisCO and its enzymatic properties. On a larger scale, the ChemCell effort supports the Subproject 4 goal of developing a general-purpose, particle-based, cell-modeling tool that can be used to model and analyze cellular response in settings where localization is important, both by this Genomics:GTL project as well as for the microbes of other Genomics:GTL projects. Ultimately, this tool could possibly be of greater use to the eukaryotic cell modeling community, where there is much more variety in cell shape and spatial localization plays an even more important role in cellular function.

This quarter we have focused on enhancing the low-level algorithms used in ChemCell to perform biochemical reactions. Our previous algorithms were heuristic rules designed to mimic the macroscopic rate constants used as ChemCell inputs. Working with our collaborator Dan Gillespie, we realized that in some reaction networks these rules can lead to inaccurate tracking of reactions and reaction cascades. We have developed a new reaction algorithm, which more closely mimics the provably accurate spatial-free stochastic simulation algorithm (SSA) pioneered by Gillespie. The new algorithm includes the effect of diffusion, thus allowing for its use in a spatial framework like ChemCell. Although the new algorithm lacks the inherent parallelism of our heuristic approaches, our hope is that it will offer a rigorous alternative for reaction networks that require its accuracy.

In FY05 we will finish the implementation of the new algorithm as an option in ChemCell and compare it to the SSA and our heuristic strategies for accuracy and speed. We plan to add a systems-biology markup language (SBML) option for ChemCell input that will make it easier to interface our tool with existing network databases and other tools. We also plan continued modeling of *Synechococcus* to study:

- 1) Competing reaction networks (light cycle, nitrate/phosphate metabolism);
- 2) Does the shape of the cell or carboxysomes or their clustering affect carbon concentration or carbon fixation;
- 3) Does the location of carbonic anhydrase inside or outside the carboxysome affect carbon fixation; and
- 4) Membrane transport through a 2-part thylakoid membrane with pores.

Inference of Protein Interaction Networks

Sandia National Laboratories

Jean-Loup Faulon, Shawn Martin, and Robert Carr

The goal of our part of Subproject 4 is to infer protein interaction networks. Predicting protein interactions is important for elucidating the functions of molecular machines and for understanding genetic and protein regulation, both of which are critical to gaining an understanding of the cell at the system level. To this end, we have developed top-down and bottom-up approaches. With the top-down approach we probe structural and dynamic properties of known or inferred biological networks, while with the bottom-up approach we predict protein-protein interactions from experimental data, including phage display and two-hybrid (such experiments are being performed within Subproject 1 for *Synechococcus*). In FY04 we completed a suite of codes designed to implement both approaches. These codes have been fully documented and are currently under review to be released as open source. We have received two requests for the bottom-up code and one request for collaboration from Dr. S. Rasheed at the University of Southern California.

In the top-down project, we presented two papers at the IEEE CSB 2004 conference at Stanford, CA. The first of these papers describes our work on our hypothesis that the power-law nature of biological networks is due to the fact that biological networks must be robust, in contrast to the currently accepted view that such networks form due to preferential attachment (Faulon, *et al.* 2004). The second of these papers describes our efforts to combine all of our tools for network inference into a pipeline with the biologist as the end-user. To test our algorithms, we have applied our methods to a yeast microarray data set and collaborated with Dr. Werner-Washburne at the University of New Mexico. Dr. Washburne is an expert in yeast genetics and has advanced two biological hypotheses based our work (Martin, *et al.* 2004a).

In the bottom-up project, we published a paper in *Bioinformatics* (Martin *et al.* 2004b) describing our method for predicting protein-protein interactions. We have also begun to apply our bottom-up code to other problems, including the problem of predicting beta strand ordering in proteins (in collaboration with Subproject 2), as well as predictions in *Synechococcus*. In the beta strand problem, we have achieved a 77 percent accuracy rate using randomly selected training and test sets from the PDB database. We are currently working on improving our results in the case of cross-family prediction. To apply our method to *Synechococcus*, where the interaction network is unknown, we have extrapolated from *Synechocystis*. We first validated this extrapolation by going from *Synechococcus* to *Nostoc*, with a 69 percent accuracy rate. Since *Synechococcus* is phylogenetically closer to *Synechocystis* than *Nostoc* is, we expect a similar accuracy for *Synechococcus*. This expectation was confirmed for carboxysome proteins when we compared our prediction with Subproject 1 newly released two-hybrid data.

Protein interaction prediction techniques such as ours, as well as experimental data such as two-hybrid, generate many false positives. This is because proteins can be found to interact in silico or in vitro, while in vivo these proteins may not interact at all, due to localization in different subcellular compartments. To reduce the rate of false positives, we plan to train a protein subcellular location predictor using a database of bacterial protein locations extracted from the literature. To improve our results in the case of cross-family prediction for beta strand binding, we also plan to base our predictions on atomistic descriptions.

To date only a few techniques have been reported in the literature to infer protein interaction networks from experimental data. These techniques are either specific to the provided data or need supplementary information, such as identification of protein domains. We have developed, validated, and published a simple and general prediction technique where the only input required is a list of protein sequences and a list of binding pairs.

Proteomic Toolshop

Oak Ridge National Laboratories

Al Geist and David Jung

In FY04 we began development of a Proteomic Toolshop codenamed *BiLab*. The Toolshop is a MATLAB®-like tool that understands and can operate on biological objects, for example, one BiLab datatype is DNA and BiLab understands the operations that can be performed on DNA, such as transcribing it into another datatype called RNA (Figure 3). The initial prototype of the Toolshop, which was just completed this month, can perform all the functions in the bioJava library and the NCBI tools library and understands a dozen biological datatypes. The Toolshop can also perform floating-point matrix operations like MATLAB. This research is part of Subproject 5. Our contribution is significant to this Genomics:GTL project and computational biology in general because the biology community needs an easy-to-use tool that can unify the biological concepts and provide a simple environment where biologists can try out new ideas. MATLAB transformed the numerical linear algebra community. Our goal is to do the same for the biological community by providing:

1. Seamless and consistent integration between command-line interaction and graphical interaction;
 - The best of both worlds and easy switching back-and-forth for the user.
2. An interactive language designed with bioinformatics in mind, rather than linear algebra;
 - Intuitive, having a shallow learning curve.
 - Full-featured to support large-scale toolbox development, that is, support for strong typing, name-space control, generic types, Object-Oriented, Design-By-Contract, etc.
 - Supports dynamic typing for interactive use and static typing for performance;
3. Allows easy integration of existing bioinformatics tools and graphical toolkits; and
4. Seamless access to all the data in online bioinformatics databases.

Our most recent accomplishments include the completion of the Toolshop language called *Scigol*, which understands biological datatypes, and the creation and debugging of the language parser, interpreter, and compiler. We created a prototype Proteomic Toolshop based on *Scigol* and integrated biological function libraries written in C, Fortran, Java, and Python. We added graphical toolkits that allow instant inline viewing of molecular structure and allow answers to be presented graphically rather than as text.

In FY05 we will make the Toolshop software more robust and portable so we can give it to researchers in this Genomics:GTL project for their use and for their feedback on the additional functions and biological datatypes required. We will use this feedback to incorporate more function libraries and datatypes. If time permits, we will investigate incorporating the protein function prediction, regulatory pathway prediction, and cell modeling applications into *Scigol* and the Proteomic Toolshop.

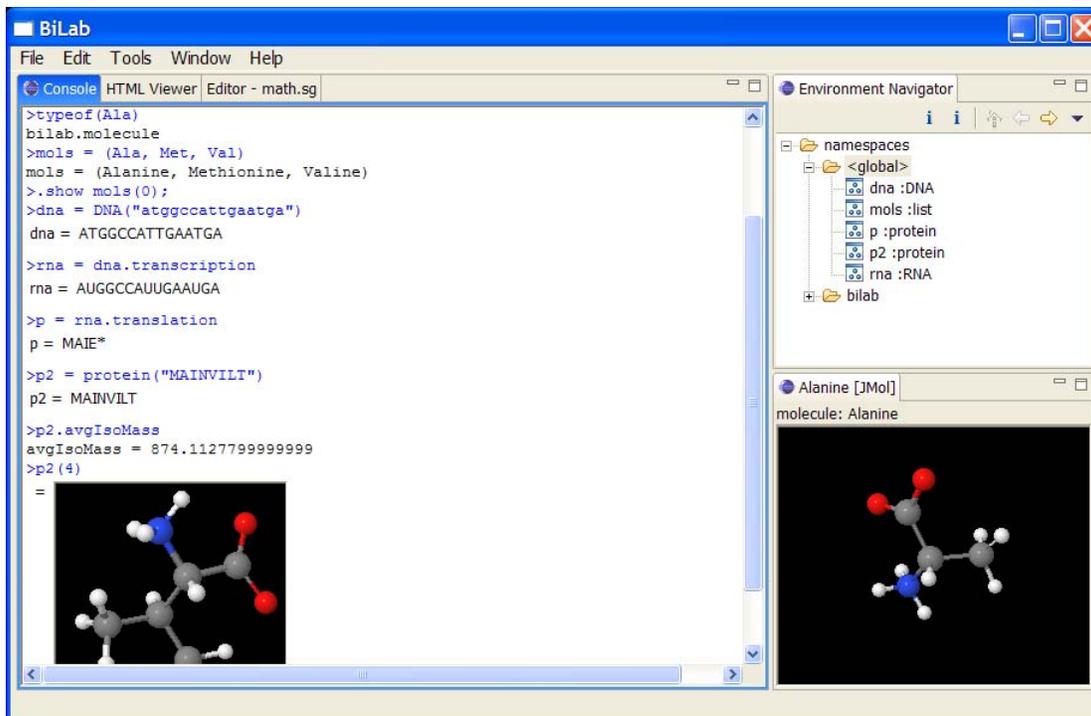


Figure 3. Screenshot of Proteomic Toolshop manipulating biological objects – DNA, RNA, proteins, and molecules

Synechococcus Encyclopedia

Oak Ridge National Laboratories

Al Geist, Nagiza Samatova, Praveen Chandramohan, and Ramya Krishnamurthy

In FY04 we developed the technology needed to construct an integrated data infrastructure that allows advanced search and queries across a large, diverse set of data sources, including sequence databases (COG, INTERPRO, SWIS PROT, TIGR, JGI, PFAM, PRODOM, SMART), structure databases (PDB, COILS, SOSUI, PROSPECT), pathway databases (KEGG), protein interaction databases (BIND, DIP, MIPS), and databases of raw mass spectrometry and microarray data from Subproject 1. A query language and integrated schema technology was developed to allow search and queries across diverse databases. We used our integrated data infrastructure to create a *Synechococcus* Encyclopedia (Figure 4) containing all currently available database knowledge about this microbe. This knowledge-base involves the integration of 23 different databases and is part of Subproject 5. Our contribution is significant to this Genomics:GTL project because the Encyclopedia contains all the experimental and analysis data generated by the other subprojects. These databases are required by the other subprojects to do protein complex and pathway predictions. Beyond data archiving, we are creating a knowledge infrastructure that provides capabilities far beyond what has been available before. The ability to construct advanced queries that require correlating and combining data from sequence annotations, protein structure, and interaction databases allows biologists to combine knowledge and see relationships that were previously obscured by the distributed nature and diverse data types in the biological databases. The technology can be used to create knowledge-bases for other organisms and in FY05 we will create an encyclopedia for *Rhodopseudomonas palustris*.

Our most recent accomplishments include the storage of the TIGR *Synechococcus* full-genome microarray data. We have added access to easy-to-use web-based analysis tools to the encyclopedia. These tools, which are being developed in the Sandia-ORNL Genomics:GTL project, include protein function characterization, protein structure prediction, comparative analysis of protein-protein interfaces, metadata entry and browsing, and electronic notebooks. Several of these tools provide transparent access to supercomputers at ORNL and around the nation. We have used the encyclopedia data and analysis tools to correctly predict the proteins making up a known membrane protein – including the proteins in the membrane itself – a task that is presently impossible by experimental mass spectrometry analysis alone.

In FY05 we will continue to support the data needs of the other SGTL subprojects. We will use the knowledge-base technology to construct encyclopedias for *Rhodopseudomonas palustris* and *Shewanella* for use by the GTL Microbial Complexes Pipeline project and Shewanella Federation, respectively. We will incorporate new pathway prediction tools developed by Subproject 3 into the *Synechococcus* Encyclopedia. This will include EXCAVATOR, CUBIC, and pMAP. If time permits, we will investigate raising the level of abstraction of the integrated data infrastructure technology to permit advanced queries across multiple organisms, for example, combining the protein structure predictions from one organism with the protein interaction predictions of another organism.

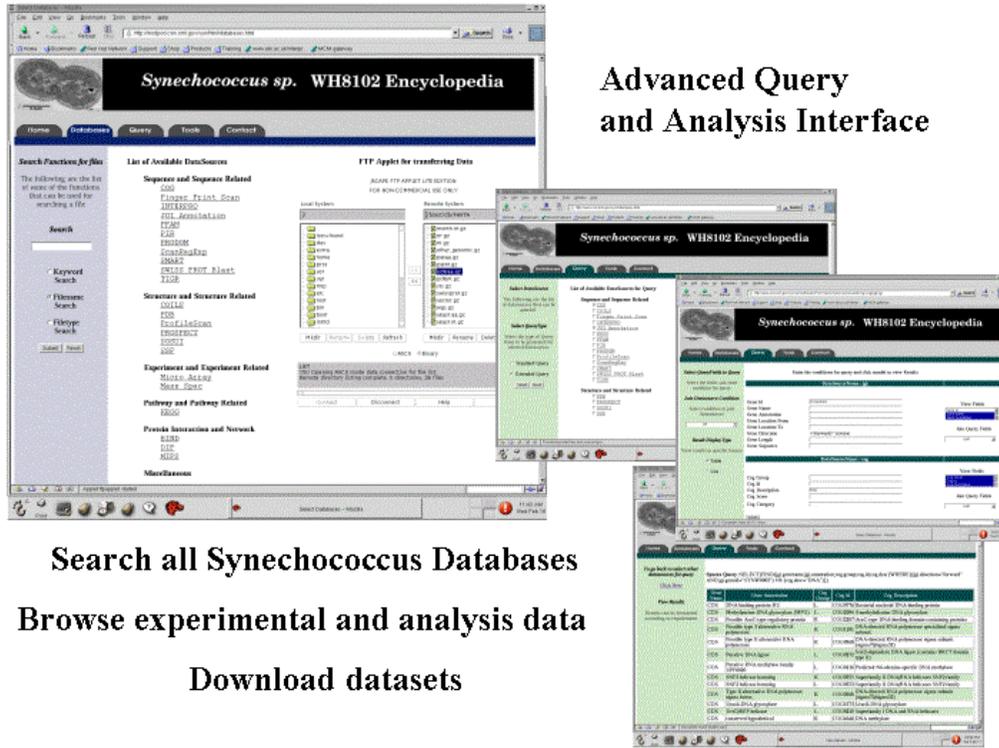


Figure 4. Synechococcus Sp. Encyclopedia

Data Entry and Browsing (DEB) Tool

Lawrence Berkeley National Laboratories

Arie Shoshani, Vijaya Natarajan, and Victor Havin

The Scientific Data Management Group at LBNL has developed a web-based Data Entry and Browsing (DEB) tool whose purpose is to facilitate capturing the metadata from experiments and laboratories and store them in a database in a computer-searchable form. The key need is to have an easy-to-use, intuitive system that integrates the metadata on all the related activities in this project. Specifically, the goal is to have a single system that captures the metadata for the Nucleotide Pool of microbes, for their microarray hybridization, and for their analysis with the hyperspectral analyzer.

The design of the DEB tool is based on inputs from the project biologists of the features that a biologist will find useful. The interface design mimics the familiar laboratory notebook format. The system is built on top of the Object-Based Database Tools (the OPM database tools developed previously at LBNL) and the data is stored in the Oracle database system. It has several important features:

- It supports multiple interrelated object-classes, such as experiment, materials, nucleotide-pool, samples, arrays, etc.;
- For each object-class, it displays a page that mimics a notebook, with pages that can be “turned” (i.e., selected by previous-next, or by number);
- Objects can be linked to each other by simple connectors, such as a “sample” object linked to its “nucleotide-pool”;
- Any file type (document, images, Excel, etc.) can be uploaded to the system and related to the metadata;
- Pages of the metadata can be printed for entry into a physical notebook – a requirement that makes sure the information is physically recorded;
- Recording a new entry can be based on a previous entry, thus avoiding the re-entry of existing entries;
- Security features to protect the metadata were developed and can be controlled by users and groups. Each experiment or related object can be assigned read-, write-, and delete-permission to other users/groups; and
- A query feature was developed to search the metadata based on conditions on the attributes of the objects.

The most powerful capability of the DEB system is that it is *schema-driven*, that is, all the interfaces to support all of the above features are generated automatically from the schema definition. Therefore, new metadata schemas can quickly be used to generate DEB interfaces as well as the underlying Oracle database for them. This feature makes this tool immediately applicable to new and/or changing databases.

The importance of an easy-to-use system for capturing metadata in Genomics:GTL cannot be overlooked, especially as an ever-growing number of experiments are conducted and a large number of datasets are collected. The ability to quickly and automatically generate metadata systems from a schema description is essential for an evolving field with multiple sources of data gathered independently. While this system is designed for this project, it can easily be applied to other Genomics:GTL efforts.

Recent accomplishments include:

- 1) The development of a version that can operate behind a firewall;
- 2) The installation of the DEB system at the Data Center at ORNL – a cooperative effort between LBNL and ORNL staff;
- 3) The addition of a feature to load a file and keep track of the location that it was loaded from;
- 4) The ability to link the metadata to files already loaded into the system, for example, large hyperspectral files; and
- 5) Metadata entries were added by SDSC into the nucleotide-pool metadatabase.

Our plans for 2005 include:

- 1) Working with the TIGR project participants to capture the microarray metadata in a format as close as possible to the MIAME standard;
- 2) Working with the Sandia project participants to capture the hyperspectral analyzer metadata and link that to the files already on the ORNL system;
- 3) Developing a DEB-lite that will remove the dependency on the OPM tools and Oracle. The goal is to package that system for use with any relational system as a backend system. This will make the tool more easily adaptable to other projects; and
- 4) Exposing other Genomics:GTL projects to this tool.

Biopathways Graph Data Manager

Lawrence Berkeley National Laboratories

Frank Olken

This effort is intended to construct a general purpose graph data management system that is capable of supporting the storage, querying, and analysis of various biopathways networks, for example, metabolic networks, signaling networks, and genetic regulatory networks, as well as other biological graph data. Such a general purpose graph DBMS could replace numerous ad hoc programs

The primary clients of this effort within the Sandia-ORNL Genomics:GTL project are the efforts to identify metabolic, signaling, and regulatory networks. Such work is being conducted by Ying Xu (University of Georgia) and Jean-Loup Faulon (Sandia). Nagiza Samatova and Michael Langston (ORNL) are working on protein interaction network annotations, for example, with clique based analyses, based on graph representations.

We also have held discussions with ORNL staff (Nagiza Samatova and Michael Langston) concerning potential collaboration on this effort. However, the collaboration has been limited by funding constraints.

The recent call for FY06 Genomics:GTL proposals calls for work in three areas:

- 1) Protein complex characterization;
- 2) Biopathways elucidation; and
- 3) Computational and data-management tools for complex biological systems.

This technology is directly relevant to all of these goals, and central to the latter two goals. Specifically, protein complexes are commonly represented as either star graphs or cliques embedded in protein interaction networks (graphs). Biopathways are the primary focus of our graph data management system. The biopathway graph representations characterize the topological structure of more detailed quantitative systems biology models and can be used as a framework on which to hang more detailed quantitative models. Finally, we note that graph databases can be used to record and query graph models of laboratory protocols and protocol execution histories, for example, for the management of large-scale biological experiments, such as proteomics, sequencing, etc.

As biopathways data are amassed, biologists are increasingly interested in doing comparative biological analyses on the biopathways graphs in a manner similar to the comparative analyses of genomic sequence data over the past decade. At present, such efforts entail the writing of ad hoc computer codes for performing the requisite graph computations and graph data management. The availability of a general-purpose graph data-management system will permit many of these comparative analyses of biopathways data to be performed by means of declarative queries, avoiding much of the special-purpose code development now required for comparative biology of pathways.

We have been working on several aspects of the BGDM:

- Refining the graph data model;
- Identifying graph queries and specifying the graph query language;

- Constructing a prototype visual query tool;
- Presenting a tutorial on graph data management at ISMB; and
- Refining the BGDM implementation architecture.

We have concluded that the graph data model should be a compound graph model, which would support overlapping hierarchical nesting of graphs, for example, hierarchically specified pathways.

We have identified several types of graph queries:

- Graph motif queries that are encoded in the query language, for example, finite graphs, regular path expressions, cycles, trees, .etc.;
- Graph operator queries, for example, subgraph isomorphism or query motif as data;
- Graph data mining queries, for example, finding frequent subgraphs; and
- Graph metric queries that compute graph statistics, such as average degree, diameter, etc.

Earlier this year I worked with Viji Natarajan (LBNL) and Kevin Keck (LBNL) on development of a prototype visual query tool. This gave us experience with visual specification of queries, including paths and practical issues of encoding queries and graphs in RDF. The prototype is not presently supported.

I presented a tutorial on graph data management at the ISMB Conference in Glasgow, UK. The half-day tutorial was well attended by approximately 143 students and led to useful conversation with other researchers in this area. We (in discussions with our colleagues at ORNL) have been working on refining the implementation architecture. We have agreement on the use of a functional query language, of a relational DBMS persistent store, and a 3-phase query processing approach that addresses initial relational selection query, main memory graph query processing, and relational join to pick up additional attributes.

In FY05 I plan to complete the specifications of a functional graph query language, initially against a simple graph data model. I also plan to further explore the implications of a compound graph model and describe proposed specifications and semantics, for example, with respect to classic graph features, graph algorithms, and graph queries. I anticipate continuing discussions with Nagiza Samatova, Michael Langston, Jean-Loup Faulon, and Ying Xu to assure that our graph data model and query language can meet their needs.

Publications and Presentations

Publications

Papers either published, in press, or submitted:

Dam, P., Z. Su, V. Olman, Y. Xu, “*In Silico* Construction of the Carbon Fixation Pathway in *Synechococcus* Sp.. WH8102,” *Journal of Biological Systems*, Invited, 2004.

Day, R., A. Borziak, A. Gorin, “PPM-Chain – De Novo Identification Program Comparable in Performance to Sequest,” *IEEE Proceedings Computational Systems Bioinformatics*, Stanford, 505-08, 2004.

Gonzales, A., Y. Light, Z. Zhang, T. Iqbal, T. Lane, A. Martino, “Proteomic Analysis of the CO₂-Concentrating Mechanism in the Open-Ocean Cyanobacteria *Synechococcus* WH8102,” *Can. J. Bot.*, Submitted, 2004.

Gorin, A., R. Day, A. Borziak, B. Strader, G. Hurst, T. Friedman, “Probability Profile Method – New Approach to Data Analysis in Tandem Mass Spectrometry,” *IEEE Proceedings Computational Systems Bioinformatics*, Stanford, 499-502, 2004.

Guo, J., K. Ellrott, W. Chung, D. Xu, S. Passovets, Y. Xu, “PROSPECT-PSPP: An Automatic Computational Pipeline for Protein Structure Prediction,” *Nucleic Acids Research*, Vol 32, W1 – W4, 2004.

Huang, J., Z. Su, Y. Xu, “The Evolution of the Phosphonate Degradation Pathways,” *Current Biology*, Submitted, 2004.

Li, Y., V. Protopopescu, A. Gorin, “Accelerated Simulated Tempering,” *Physics Letters A*, 328(4-5), 274-283, 2004.

Li, G., X. Wang, X. Qi, B. Zhu, Y. Xu, “A Linear-time Algorithm for Computing the Translocation Distance Between Signed Genomes,” *Theoretical Computer Science*, in press, 2004.

Liu, Z., D. Chen, H. Bensmail, Y. Xu, “Gene Expression Data Clustering with Kernel Principal Component Analysis,” *Journal of Bioinformatics and Computational Biology*, in press, 2004.

Liu, Z., D. Chen, Y. Xu, “Logistic Support Vector Machines and Their Application to Gene Expression Data,” *IEEE Transaction on Computational Biology and Bioinformatics*, Submitted, 2004.

Liu, Z., F. Mao, J. Guo, B. Yan, Y. Xu, “Quantitative Validation of Protein-DNA Interaction in Transcription Process: Distance-dependent Knowledge-based Potential Considering Multi-body Interaction,” *JBC*, Submitted, 2004.

Mao, F., P. Dam, V. Olman, Z. Su, Y. Xu, “An Integer Programming Algorithm for Mapping of Microbial Pathways and Regulons Across Multiple Genomes,” Submitted, 2004.

- Martin, S., D. Roe, J.-L. Faulon, "Predicting Protein-Protein Interactions Using Signature Products," *Bioinformatics*, in press, 2004.
- Shah, M., S. Passovets, D. Kim, K. Ellrott, L. Wang, I. Vokler, P. LoCascio, D. Xu, Y. Xu, "A Computational Pipeline for Protein Structure Prediction and Analysis at Genome Scale," *Bioinformatics*, 19, 1985 – 1996, 2003.
- Stork, C., M. Keenan, D. Haaland, "Multivariate Curve Resolution for the Analysis of Remotely Sensed Thermal Infrared Hyperspectral Images," *Proceedings of SPIE*, 2004.
- Su, Z., P. Dam, V. Olman, B. Palenik, Y. Xu, "Characterization of Nitrogen Assimilation Pathways in *Synechococcus* Sp. WH8102 Through Computational Prediction and Experimental Validation," Submitted, 2004.
- Su, Z., P. Dam, F. Mao, X. Chen, V. Olman, T. Jiang, B. Palenik, Y. Xu, "Towards Computational Inference of Regulatory Pathways in Prokaryotes: An application to Phosphorus Assimilation Pathways in *Synechococcus* Sp. WH8102," *Genomes Research*, Submitted, 2004.
- Timlin, J., D. Haaland, M. Sinclair, M.J. Martinez, M. Werner-Washburne, "Hyperspectral Microarray Scanning: Impact on the Accuracy and Reliability of Genomic Data," *Genome Biology*, Submitted, 2004.
- Xu, D., P. Dam, D. Kim, M. Shah, E. Uberbacher, Y. Xu, "Characterization of Protein Structure and Function at Genome-scale using a Computational Prediction Pipeline," *Genetic Engineering: Methods and Principles*, 269-293, 2003.
- Xu, D., V. Olman, L. Wang, Y. Xu, "EXCAVATOR: a Computer Program for Gene Expression Data Analysis," *Nuclear Acid Research*, 31(19), 5582-5589, 2003.
- Yan, B., C. Pan, V. Olman, B. Heittich, Y. Xu, "A Graph-theoretic Approach to Separation of b- and y-ions in Tandem Mass Spectra," *Bioinformatics*, in press, 2004.
- Yan, B., Y. Qu, F. Mao, V. Olman, Y. Xu, "PRIME: A Mass Spectrum Data Mining Tool for *De Novo* Sequencing and PTMs Identification," *Journal of Computer Science and Technology*, in press, 2004.
- Yip, G., E. Zuiderweg, "A Phase Cycle Scheme That Significantly Suppresses Offset-Dependent Artifacts in the R(2)-CPMG (15)N Relaxation Experiment," *J. Magn Reson.*, 171(1):25-36, 2004.

Presentations

- Chen, X., Z. Su, Y. Xu, T. Jiang, "Computational Prediction of Operons in *Synechococcus* Sp. WH8102," *Proceedings of the 15th International Conference on Genome Informatics*, 2004.
- Chen, Y., T. Joshi, Y. Xu, D. Xu, "Towards Automated Derivation of Biological Pathways Using High-Throughput Biological Data," *Proceedings of the IEEE Conference on Bioinformatics and Biotechnology*, 2003.
- Dam, P., Z. Su, V. Olman, Y. Xu, "Computational Reconstruction of the Carbon Fixation Pathway in *Synechococcus* Sp. WH8102," *Proceedings of International Conference on Bioinformatics and its Applications*, 2004.

- Dam, P., Z. Su, X. Chen, V. Olman, T. Jiang Y. Xu, "In silico Construction of the Carbon Fixation Pathways in *Synechococcus* Sp. WH8102," *Proceedings of the Third IEEE Computer Science Society Conference on Bioinformatics*, 2004.
- Faulon, J.-L., S. Martin, R. Carr, "Dynamical Robustness in Gene Regulatory Networks," *Proceedings IEEE-CSB04*, vol. 3, 626-627, 2004.
- Gonzales, A., Y. Light, Z. Zhang, T. Iqbal, T. Lane, A. Martino, "Proteomic Analysis of the *Synechococcus* WH8102 CCM with Varying CO₂ Concentrations," *12th International Meeting of Microbial Genomes*, Lake Arrowhead, CA, 2004.
- Gonzales, A., Y. Light, Z. Zhang, T. Iqbal, T. Lane, A. Martino, "Proteomic Analysis of the *Synechococcus* WH8102 CCM with Varying CO₂ Concentrations," *5th International Symposium on Inorganic Carbon Utilization by Aquatic Photosynthetic Organisms*, Montreal, Canada, 2004.
- Haaland, D., J. Timlin, M. Keenan, H. Jones, C. Stork, D. Melgaard, M. Sinclair, "Multivariate Analysis of Hyperspectral Images for Biotechnology Applications," *Meeting of the American Chemical Society*, Invited, Philadelphia, PA, 2004.
- Haaland, D., J. Timlin, H. Jones, M. Keenan, C. Stork, M. Sinclair, "Multivariate Analysis of Hyperspectral Images for Biotechnology Applications," *9th Conference on Chemometrics in Analytical Chemistry*, Invited, Lisbon, Portugal, 2004.
- Haaland, D., J. Timlin, M. Sinclair, M. Keenan, M. Van Benthem, M. Martinez, A. Aragon, G. Quinones, M. Werner-Washburne, "Multivariate Analysis of Hyperspectral Images for Biotechnology Applications," *Chips to Hits Conference*, Invited, Boston, MA, 2004.
- Haaland, D., J. Timlin, M. Keenan, H. Jones, C. Stork, D. Melgaard, M. Sinclair, "New Approaches for Understanding Multivariate Curve Resolution Applied to Hyperspectral Images," *Federation of Analytical Chemistry and Spectroscopy Societies Meeting*, Invited, Portland, OR, 2004.
- Jones, H., D. Haaland, J. Timlin, E. Thomas, "Understanding the Limitations of Multivariate Curve Resolution when Applied to Hyperspectral Images," *Federation of Analytical Chemistry and Spectroscopy Societies Meeting*, Portland, OR, 2004.
- Mao, F., Z. Su, V. Olman, D. Chung, Y. Xu, "Pathway Mapping With Operon Information: An Integer-Programming Method," *Proceedings of the Third IEEE Computer Science Society Conference on Bioinformatics*, 2004.
- Martin, S., G. Davidson, E. May, J.-L. Faulon, M. Werner-Washburne, "Inferring Genetic Networks from Microarray Data," *Proceedings IEEE-CSB04*, vol. 3, 566-569, 2004.
- Olman, V., D. Xu, Y. Xu, "A New Framework for Biological Data Clustering using Minimum Spanning Trees," *Proceedings of The 7th Pacific Symposium on Biocomputing*, 2003.
- Olman, V., H. Peng, Z. Su, Y. Xu, "Mapping of Microbial Pathways Through Constrained Mapping of Orthologous," *Proceedings of The IEEE Computational Systems Bioinformatics Conference*, 2004.

- Plimpton, S., A. Slepoy, S. Means, M. Rintoul, "Modeling Cellular Response," *PNNL Northwest Symposium for Systems Biology*, Richland, WA, 2004.
- Plimpton, S., A. Slepoy, "ChemCell: a Particle-based Cell Model of Protein Interactions," *PSC Workshop on Computational Methods for Spatially Realistic Cellular Simulations*, Pittsburgh, PA, 2004.
- Stork, C., M. Keenan, D. Haaland, "Multivariate Curve Resolution for the Analysis of Remotely Sensed Thermal Infrared Hyperspectral Images," *SPIE Annual Meeting*, 2004.
- Su, Z., A. Dam, X. Chen, V. Olman, T. Jiang, B. Palenik, Y. Xu, "Computational Inference of Regulatory Pathways in Microbes: an Application to the Construction of Phosphorus Assimilation Pathways in *Synechococcus* WH8102," *Proceedings of the 14th International Conference on Genome Informatics*, Tokyo, Japan, 2003.
- Su, Z., P. Dam, X. Chen, V. Olman, T. Jiang Y. Xu, "Computational Construction of Nitrogen Assimilation Pathway in Cyanobacteria *Synechococcus* Sp. WH8102," *Proceedings of the Third IEEE Computer Science Society Conference on Bioinformatics*, 2004.
- Timlin, J., D. Haaland, M. Sinclair, "A Hyperspectral Imaging System with Multivariate Data Analysis for Increased Throughput," *118th AOAC Annual Meeting & Exposition*, Invited, St. Louis, MO, 2004.
- Timlin, J., D. Haaland, M. Sinclair, "Hyperspectral Imaging for Endogenous and Exogenous Fluorophore Differentiation in Live Cells," *Federation of Analytical Chemistry and Spectroscopy Societies Meeting*, Portland, OR, 2004.
- Wang, P., Z. Su, Y Xu, "A Knowledge Base for Computational Pathway Reconstruction in *Synechococcus* Sp. WH8102," *Proceedings of the Third IEEE Computer Science Society Conference on Bioinformatics*, 2004.
- Yan, B., C. Pan, R. Heittich, V. Olman, Y. Xu, "Separation of Ion Types in Tandem Mass Spectrometry Data Interpretation –a Graph-theoretic Approach," *Proceedings of The IEEE Computational Systems Bioinformatics Conference*, 2004.

*This work was funded in part or in full by the U.S. Department of Energy's Genomes to Life program (www.doegenomestolife.org) under project, "Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org).*

DISTRIBUTION FOR OCTOBER 2004 GTL QUARTERLY SAND REPORT

SAND2005-

10	MS-0885	Grant Heffelfinger, 1802
1	MS-9018	Central Technical Files, 8945-1
2	MS-0899	Technical Library, 9616