SAND REPORT

SAND2004-5762
Unlimited Release
Printed January 2005

Genomes to Life Project Quarterly Report June 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.

Sandia National Laboratories

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

SAND2004-5762
Unlimited Release
Printed January 2005

Genomes to Life Project Quarterly Report June 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 June 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 CO_2 are important terms in the global environmental response to anthropogenic atmospheric inputs of CO_2 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

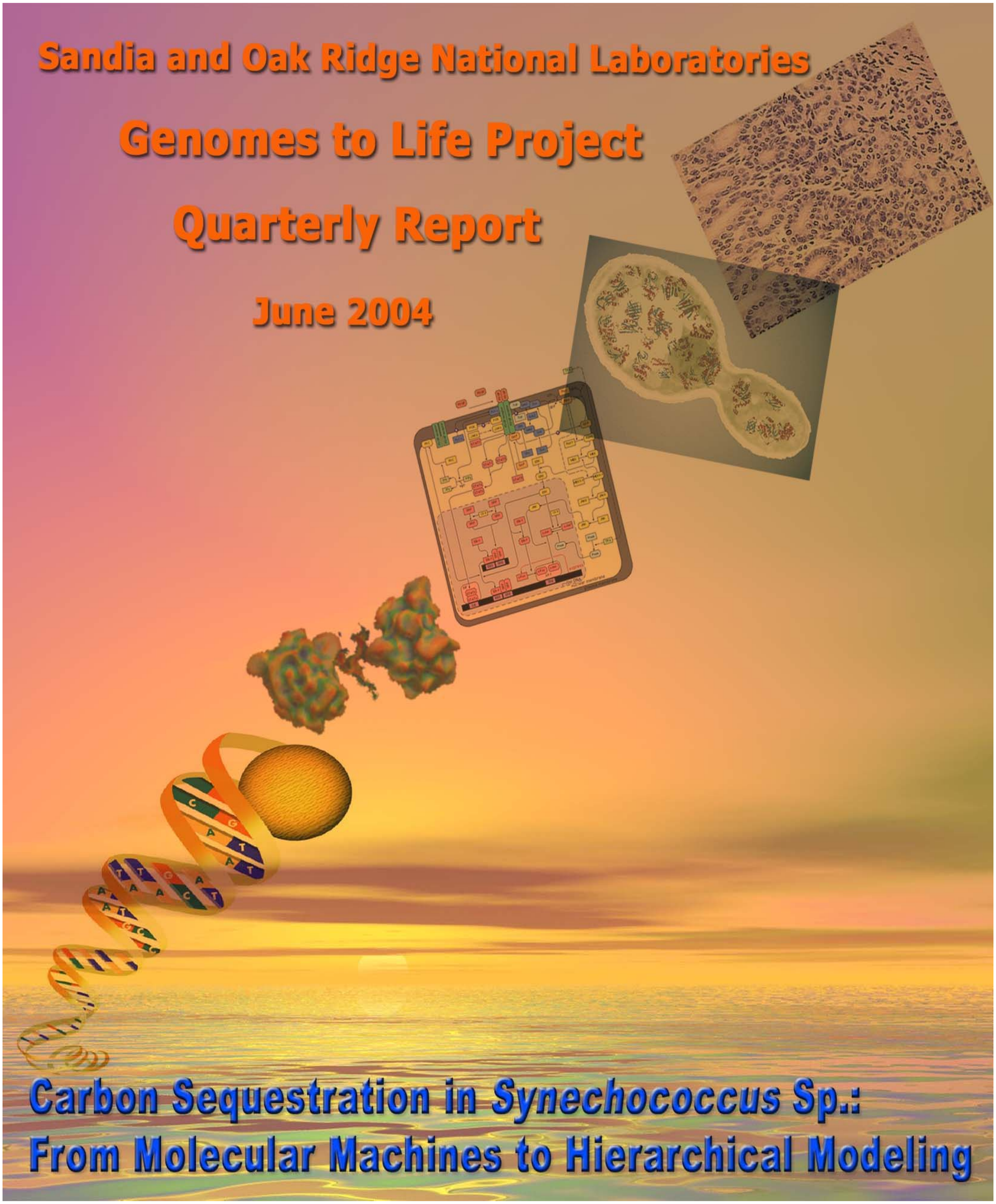
Acknowledgment

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

Grant S. Heffelfinger[1]*, Anthony Martino[2], Andrey Gorin[3], Ying Xu[10,3], Mark D. Rintoul[1], Al Geist[3], Matthew S. Ennis[1], Hashimi M. Al-Hashimi[8], Andrea Belgrano[12] Paul Crozier[1], George S. Davidson[1], Jean Loup Faulon[2], Laura J. Frink[1], Damien Gessler[12], David M. Haaland[1], William E. Hart[1], Erik Jakobsson[7], Tao Jiang[9], Howland Jones[1], Todd Lane[2], Shawn Martin[1], Shawn Means[1], Thomas Naughton[3], Frank Olken[4], Brian Palenik[6], Hoony Park[3], Steven J. Plimpton[1], Susan Rempe[1], Diana C. Roe[2], Nagiza F. Samatova[3], Arie Shoshani[4], Alex Slepoy[1], Charlie E. M. Strauss[5], Zhengchang Su[10], Edward V.Thomas[1], Jerilyn A. Timlin[1], Xiufeng Wan[11], Dong Xu[11], Zhaoduo Zhang[2]

*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

June 2004

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

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

Executive Summary

Within our experimental efforts, we continue the analysis of the composition and protein-protein interactions present in the carboxysome of *Synechococcus* Sp. with the development of optimized isolations and interaction assays. Intact carboxysomes have been isolated. A number of protein interactions within the carboxysome have been determined by 2-hybrid assays. Hyperspectral imaging and multivariate data analysis has been completed on several microarrays enabling both improved quality control and increased throughput.

Our computational discovery work has resulted in new algorithms and programs released. We released the first version of PPM-Charger – our novel algorithm for parent-ion charge determination in tandem mass spectrometry. We also developed a new computational procedure that predicts clusters of specificity-determining residues among highly homologous functional protein groups. Our long time-scale molecular dynamics simulations of a RuBisCO complex suggest that, contrary to the Schlitter/Wildner theory, gaseous ligands can escape with the trap door closed. In addition, we successfully predicted the most-stable protonation state and established a rate for one of the proton transfer reactions of RuBisCO.

Our work on computational methods characterizing *Synechococcus* Sp. regulatory pathways has led to our continued development of the computational pipeline for microbial pathway inference and reconstruction. We are in the process of completing a number of method developments in protein-protein interaction predictions, operon structure prediction, transcription-factor binding-site identification, mapping of known biological pathways, statistical analysis of microarray gene expression data, and automated literature mining.

Our computational systems biology efforts have allowed us to continue development and testing of our particle-based ChemCell model of protein/protein interactions within cells. Within our top-down approach, we have used our inference and analysis network tools to functionally characterize eight organisms, including six bacteria. We find that functional distances between the studied organisms match their phylogenetic distances well. We also observe that there are binding rules between protein functions, that is, most proteins bind to proteins having the same function, while interfunction bindings occur rarely. Related to our bottom-up approach, the precision versus sensitivity of our signature-product method compares favorably with the technique recently published by Jansen, et al. (2003). Finally, we have continued to refine our linkage between hierarchical levels and are now testing component levels against the data, such as that available from the HOT (Hawaiian Ocean Time-series) program.

Our computational work environments effort made significant progress this quarter toward establishing a data and computational infrastructure for use by the project. The *Synechococcus* Encyclopedia became available for all project members to use. New search capabilities were added to the Encyclopedia based on user feedback. Work continued on the MATLAB^{R}-like biology tool, “BiLab.”

Subproject 1: Experimental Elucidation of Molecular Machines and Regulatory Networks in *Synechococcus* Sp.

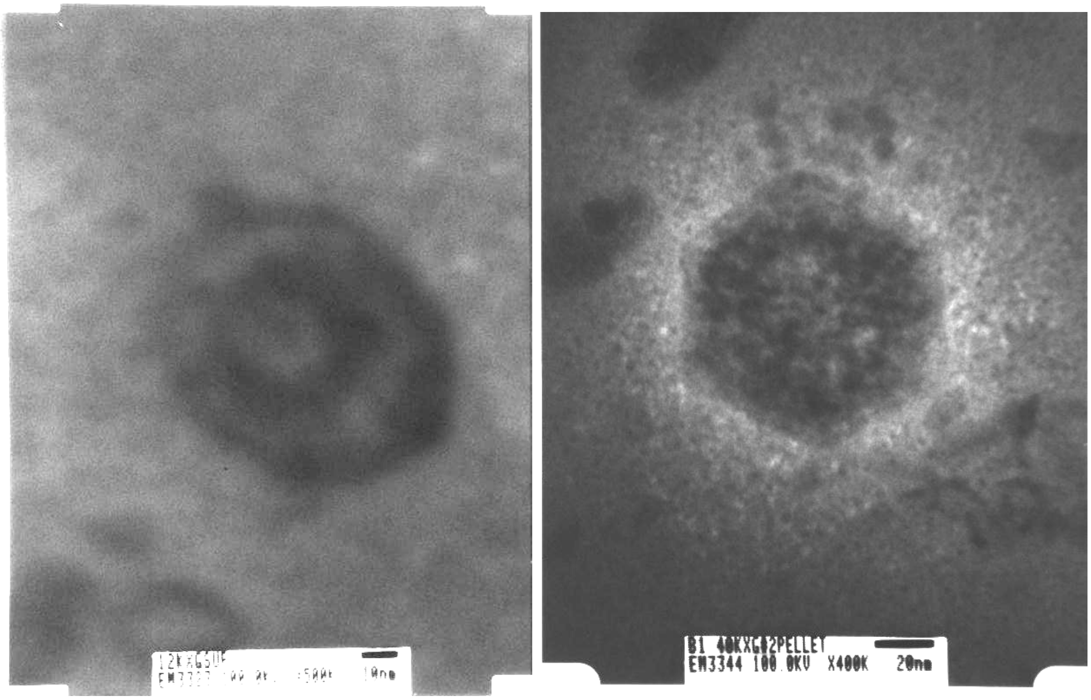


Figure 1-1. Electron micrographs of *Synechococcus* Sp. WH8102 carboxysomes show maturation of organelles. Underdeveloped carboxysomes appear ringed with incomplete shell structures. Fully developed organelles show distinct hexagonal forms with sharp edges and granular interiors.

We continued the analysis of the composition and protein-protein interactions present in the carboxysome with the development of optimized isolation techniques and a number of interaction assays. Intact carboxysomes have been isolated and show a distinct hexagonal shape with a granular interior (Figure 1.1). We started sucrose gradient purification of isolated carboxysomes this quarter. Mass spectrometry proteomic analysis of carboxysome fractions still indicates some contamination. Protein interactions within the carboxysome are being studied using 2-hybrid assays, phage display, solution NMR methods, and protein affinity-tagged pull down experiments. A number of protein interactions within carboxysomes have been determined using 2-hybrid assays. Additionally, we are continuing with knock-out and native expression of tagged carboxysome gene strategies. ccmk2 knock-outs are not viable under ambient CO_2 concentrations, so we are attempting high CO_2 concentrations. Also, this quarter we have continued a set of experiments to measure the gene expression profiles using microarray technology of phosphate-regulatory mutants under phosphate-depleted conditions and different nitrogen sources. RNA isolation, DNA hybridization to gene chips, and data analysis continues. Hyperspectral imaging and multivariate data analysis has been completed on several microarrays to improve quality control and to increase microarray throughput through use of multiple labels.

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).

Accomplishments

Carboxysome Purification and Composition Analysis: In the past, we have described the development of a protocol using combinations of low- and high-speed centrifugation steps separated by the addition of the divalent $MgSO_4$ ion, Triton X-100, and YPERS^{TM} (yeast gentle lysis buffer) to purify carboxysomes. We continue to use SDS-PAGE, mass spectrometry, and transmission electron microscopy to analyze the purification of carboxysomes. We reported in the past that electron microscopy results indicated a large fraction of the carboxysomes were being broken. In the last quarter, we attempted the purification at lower lysis pressures (10K psi versus 20K psi). We now observe intact carboxysomes with distinct hexagonal shapes and well-defined edges (Figure 1.1). The interiors are granulated consistent with the presence of high concentrations of RuBisCO.

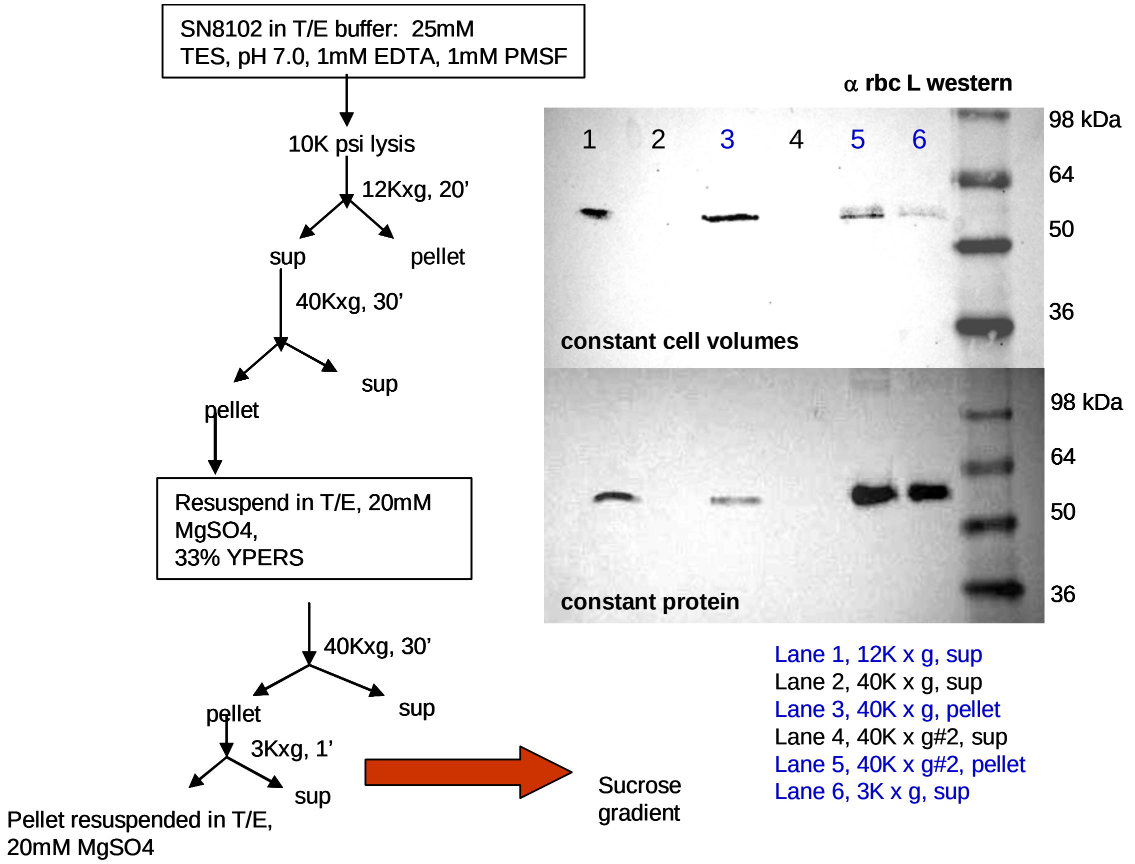


Figure 1-2. Schematic of carboxysome purification protocol with characterization of steps along the way by RuBisCO western blotting. rbcL westerns normalized for constant cell volumes indicate RuBisCO is found exclusively in the particulate fractions. rbcL westerns normalized for constant protein loading show enrichment of RuBisCO in the particulate fractions. Comassie stains of the gels, however, show many common bands across particulate and soluble fractions, indicating incomplete purification. Mass spectrometry proteomics results agree.

SDS-PAGE and western characterization of the purification protocol indicates RuBisCO, and therefore carboxysomes, are exclusively present in the particulate (pellet) fractions and that the components are enriched with further purification steps (Figure 1-2). Comassie stains of the gels, however, show many common bands across particulate and soluble fractions indicating incomplete purification, and mass spectrometry proteomics results agree. We are now attempting to purify carboxysomes from the particulate fraction using sucrose gradients and percoll-sucrose gradients.

Last quarter, we described the synthesis of antibodies to components of the carbon concentrating mechanism (CCM) in *Synechococcus* 8102. We now have antibodies to carboxysome shell proteins ccmK and csoS2 and two putative carbonic anhydrases. Western characterization of the purification protocol using the csoS2 antibody agrees with RuBisCO blotting (Figure 1-3). The shell protein exists only in the particulate fraction. ccmK is of the same size as an antibody cross-reactivity species, and therefore no information can be taken from those westerns.

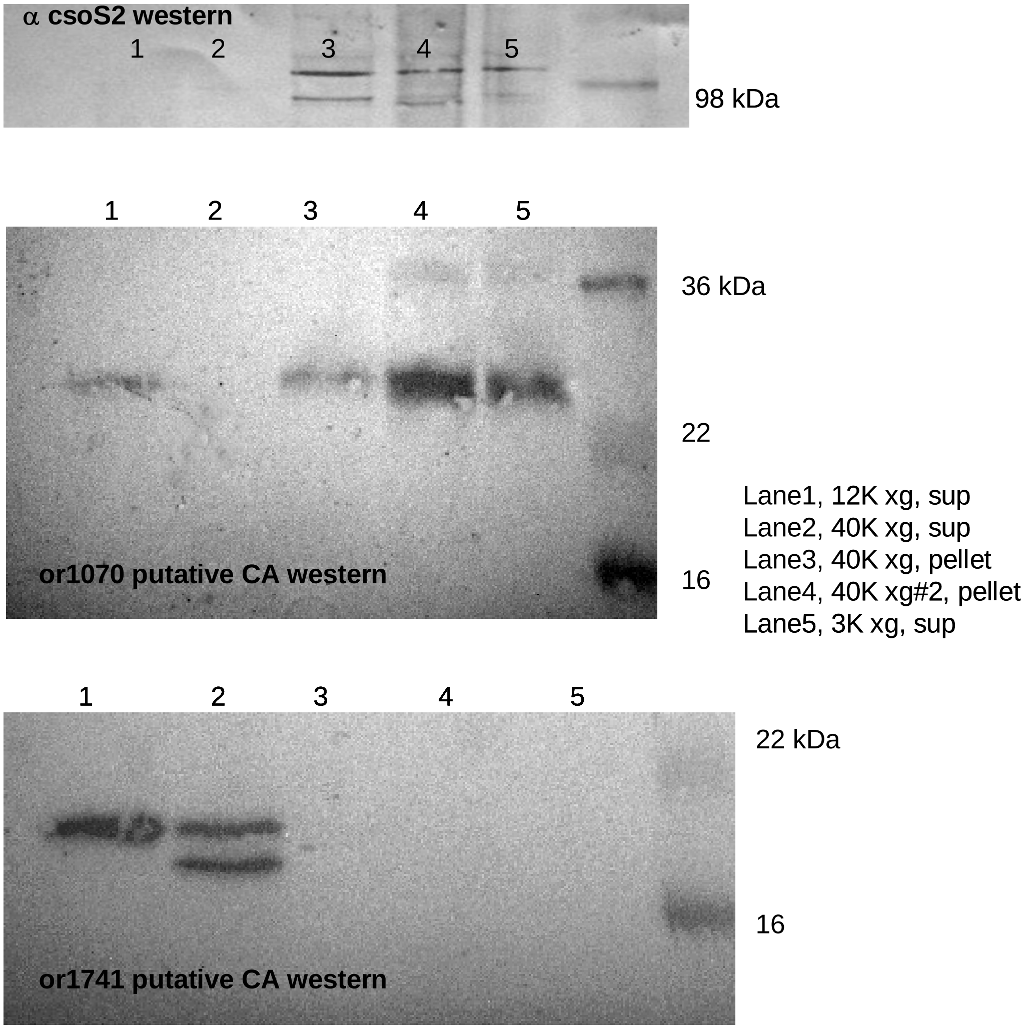


Figure 1-3. Western blots of the carboxysome shell protein csoS2 and two putative carbonic anhydrases (CA). csoS2 results mimic RuBisCO results and are consistent with the presence of the carboxysome in the particulate phase. Two putative CAs show different results. or1070 separates with the particulate fraction and may be present in the carboxysome. or1741 appears to be soluble.

Westerns of two putative carbonic anhydrases show interesting results. or1070 separates with the particulate fraction and or1741 appears soluble. This probably indicates functional distinction between the two. Further characterization is required.

Interactions within the Carboxysome: Several assays to determine interactions within the carboxysome including 2-hybrid, phage display, NMR methods, and protein affinity pull down experiments are under way. In the last quarter, we have detected numerous pair-wise interactions between components of the carboxysomes using bacterial 2-hybrid analysis (Figure 1-4).

Cotransformation of pBT and pTRG carrying cso operon genes
Cotransformation matrix

	pBT- rbcL	csoS2	csoS3	pepA	pepB	rbcS	pBT only
pTRG- rbcL	++ 4	○	○	○	○	○	
rbcS	++ 4 ○	4 ○	4 ○	4 ○	4 ○	○	
ccmk1	4 ○	4	4	4	4	+ ☆	
ccmk2	4	○	☆	+ ☆	☆	+ ☆	
csoS2	+++ 4 ○	+++ 4 ○	4	+++ 4 ○	++ ☆	++ ☆	○
csoS3	4	+ ☆	☆	+ ☆	☆	☆	
pepA	+++ 4 ○	++ ○	☆		+ ☆	+ ☆	
pepB	4	○					
pTRG- only	○						○

Figure 1-4. Interaction matrix of cso operon genes analyzed by bacterial 2-hybrid screening. Positive interactions have been reproducibly detected and analysis is ongoing.

Positive control: pBT-LGF2/ pTRG-GAL11 (+++++).
Negative controls: 1. No Plasmid; 2. Empty vectors only; 3. pTRG-csoS2/pBT only; 4. pBT-rbcL/pTRG only.
Symbols: 1. 4 cotransformations done on 03/23/04; 2. ○ cotransformation done on 04/07/04; 3. ☆ cotransformation done on 05/20/04; 4. + indication of number of colonies.

Some tests have been done multiple times and testing continues, with proper controls always tested. Analysis is ongoing, but we are highlighting one result now. The experiments indicate that csoS2 in pTRG interacts with rbcL, rbcS, and pepB. These interactions are not observed when csoS2 is in pBT. Therefore, the interactions are either directionally dependent or are false positives. csoS2 shows positive interactions with pepA in either vector. This is viewed as a highly likely interaction. Complete matrix analysis will be presented when the experiments are complete.

Phage display is currently being used to test interactions within the carboxysome. rbcL and rbcS are being synthesized as substrates and 7-mer and 12-mer peptide libraries are being screened. Cloning of rbcL and rbcS with His tags was described last quarter. In the past, we have described our efforts to purify rbcS. We now believe our protocol is optimized, and we are synthesizing purified rbcS (Figure 1-5).

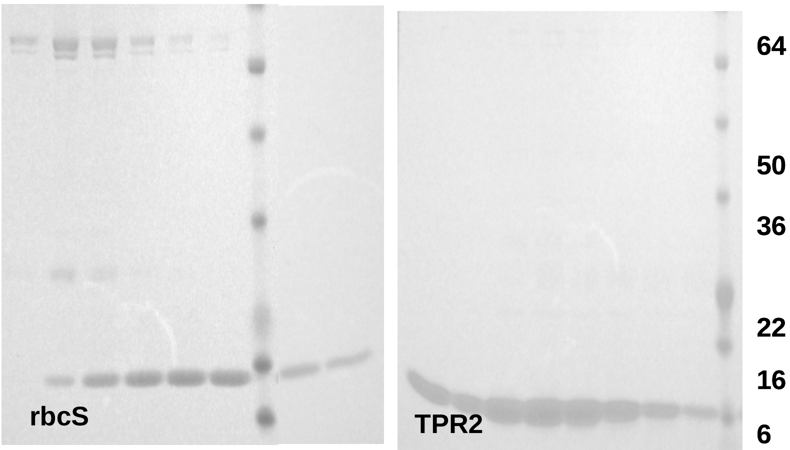


Figure 1-5. Purification efforts for rbcS and TPR2 as substrates in phage-display experiments.

Protein Binding Domains and Interaction Networks: In the past, we completed initial control phage-display experiments with 7-mer and 12-mer libraries and cloned three TPR repeat units from a single PSA E protein. In the last quarter, we attempted to purify the TPR repeat unit peptides for phase-display analysis. We now believe our protocol is optimized and we are synthesizing purified TPR domains (Figure 1-5).

Carboxysome Interactions via NMR Studies: We are focusing on the development and application of experimental solution NMR methods for high-throughput structural and dynamic characterization of protein-protein interactions. For the past few months, the University of Michigan team has been working on the purification and isolation of the small subunit of RuBisCO (rbcS) as the chosen primary NMR target for investigation. The preparation and purification of NMR quantities of rbcS is necessary before we can begin to apply the various NMR methods that we have developed for probing protein-protein interactions involving rbcS. These methods include automatic resonance assignment, developed in collaboration with the Gorin group at ORNL, and NMR spin-relaxation-based characterization of binding sites, developed in collaboration with the Zuiderweg group (University of Michigan).

We have received plasmids for both the large and small subunits of the enzyme from Zhao Zhang in the Martino laboratory. In previous work, Zhao was getting excellent expression, as evidenced by the heavy band at approximately the correct molecular weight. However, the protein corresponding to this band was not binding to the nickel-NTA resin. We reasoned that this may occur because the enzyme was being cleaved near the N-terminus where the His tag is localized before it could be purified. In addition, from the crystal structure, the N-terminus of the rbcS appears to be unstructured, so the His tag should be accessible for binding the nickel resin (Figure 1-6). Therefore, we chose to concentrate on inhibiting the potential cleavage problem.

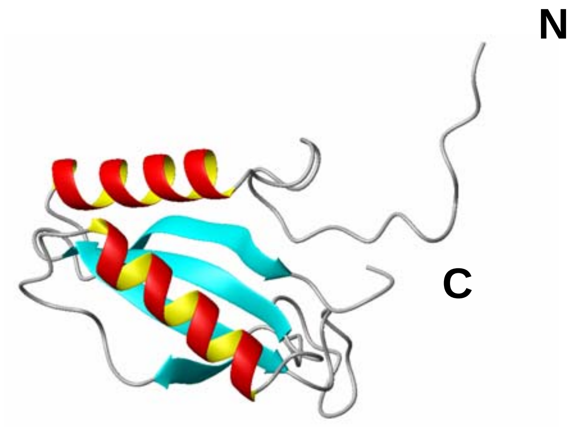


Figure 1-6. Ribbon diagram of the small subunit of RuBisCO (1EJ7).

In an attempt to alleviate this problem, protease inhibitors were added in the hope of preventing the cleavage. Four 3-mL cultures were grown to test for expression and all appeared to provide good expression. The culture was then scaled-up to 50 mL in duplicate. Just before sonication, either PMSF or PMSF + protease cocktail (Complete Mini, EDTA-free, Roche) was added. The supernatant was applied to the nickel-NTA column in standard buffer (50 mM Tris, pH 8.0, 300 mM NaCl, 5 mM β-mercaptoethanol) and eluted with 250 mM imidazole added to the same buffer. In both cases, essentially all of the rbcS was bound to the column (Figure 1-7).

With the addition of the protease inhibitors, the protein seems to be intact. However, after the nickel-NTA column, the rbcS was only approximately 50% pure. The protein was then applied to a Q-sepharose column during this step, so it is imperative to maintain the protease inhibitors throughout the preparation. Another small-scale purification yielded enough protein to attempt a 1-D NMR spectrum of the unlabeled sample. Unfortunately, the buffer conditions that were chosen were not compatible with the protein and it precipitated before the spectrum could be measured.

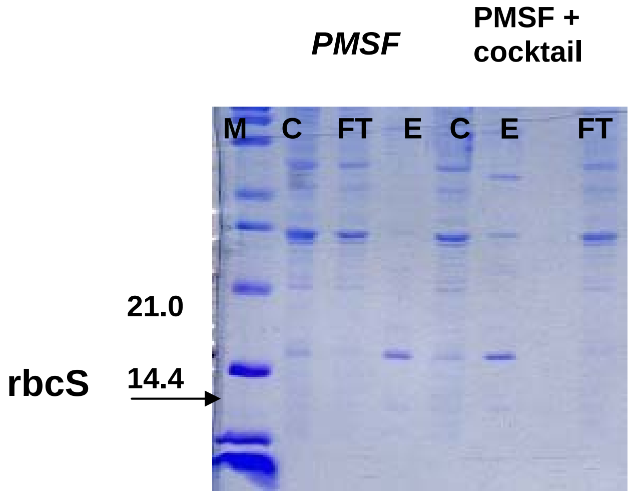


Figure 1-7. SDS-PAGE of rbcS purification by nickel-NTA column. **M**- molecular weight markers, **C**- cell lysate, **FT**- column flow-through, **E**- eluant from nickel column.

We have obtained the equipment to purify the protein on a larger scale and are currently working on the purification of a 1-L preparation of rbcS. This should allow the purification protocol to be finalized and provide enough protein for optimization of buffer conditions for the NMR experiments. Once the purification process is established, the rbcS will be 15N-labeled for an initial NMR screen and optimization of alignment media. Specifically, we will employ the array of buffer conditions that were chosen to sample a range of ionic strength, pH as well as charged buffers, in the previous report and examine both 1-D and 2-D 15N -1H HSQC spectra. We have also initiated preparation and purification of the rbcL subunit, which will allow measurement of the rbcS activity as a further check of its proper folding in solution.

Gene Regulatory Network by Microarray Analysis: We have continued a set of experiments to measure the gene expression profiles using microarray technology of phosphate-regulatory mutants under phosphate-depleted conditions and different nitrogen sources. RNA isolation, DNA hybridization to gene chips, and data analysis continues.

The SNL statistically designed, full genome *Synechococcus* microarrays from TIGR have been delivered to SNL. Data analysis has begun on this first set of hybridized arrays and will be completed in the next quarter. We have performed hyperspectral scans on three slides exhibiting minor irregularities: green background smears, high background, red speckle, and possibly green contaminant. Initial multivariate data analysis has shown the red speckle to be Cy5 dye. The high backgrounds can also be attributed to Cy3 and Cy5 dyes. There is initial evidence of a green contaminant present at very low levels in the buffer-only printed spots. Due to the discovery of artifact present in that data as a result of the rotation of the CCD detector relative to the spectrometer image-plane and the extremely weak and overlapping nature of this contaminant, these scans will be repeated to verify contaminant presence and significance following the completion of the alignment adjustments to remove the artifact.

We have completed the majority of the optimizations of our hyperspectral scanner and the scanner has been moved to a new location to accommodate its growing applications. The performance of the new positioners has been characterized. This new hardware provides a factor of five improvement in frame-to-frame variability over the previous positioners. We have continued to improve the acquisition and analysis software. Most of the recent work has been focused on tools to improve and automate hardware alignment.

Pigment Characterization by Hyperspectral Imaging: We have conducted initial experiments to build collaboration with Wim Vermaas's group at Arizona State University. Their group has ongoing efforts to understand cyanobacteria, specifically *Synechocystis*, focusing on reactions involving photosynthesis, respiration, and carbon fixation. They have developed many deletion mutants lacking a specific gene and are using these to assemble a blueprint of the organism's pathways. One primary obstacle has been in resolving the highly overlapping fluorescence emissions of the native pigments (phycobilisomes and chlorophylls) in the bacterium, an area where hyperspectral imaging could be extremely useful. We have obtained samples from the ASU lab of wild type *Synechocystis* Sp. PCC 6803 and two gene variants (*ch*L$^-$ and PS I-less/ *ch*L$^-$) and have performed necessary modifications to our scanner to image these small samples at 0.83-μm spatial resolution. Each of the *Synechocystis* samples were resuspended individually in media and hyperspectral images were acquired while the cells were in a live state. Initial multivariate data analysis indicates that each strain of cells has unique and complex spectral features. Although a full analysis, including spectral assignment, has not yet been completed, these initial results were consistent with Wim Vermaas's expectations and provided the group with spectral information previously unavailable. Future experiments are planned to optimize experimental parameters and to increase our understanding of the emission of these native pigments. A proposal with ASU was submitted to DOE/ASCR to take this collaboration to a higher level.

Progress Towards Milestones

FY03

Aim 1. Establish *Synechococcus* cultures (11/02). Complete.

Aim 3. PCR amplify genes for substrate binding proteins. Express in E. coli (11/02). Three TPR repeats and rbcL and rbcS have been cloned, expressed, and purified. Further cloning is in process.

Aim 3 . Construct improved hyperspectral scanner (parts purchased in 4th quarter of FY02). Quantify improvement in accuracy and dynamic range of new scanner (12.02). We have characterized the motion stages on our improved scanner and determined that they impart a factor of 5 reduction in data variability over the previous motion platform. We have also begun development of software tools to assist in automated alignment of the instrument.

Aim 2. Expression and purification of 15N-, and 15N/13C-, and 15N/13C/2H- isotopically enriched proteins (1/03). Clones and protocols produced at Sandia were shipped to the University of Michigan. UM has successfully synthesized purified rbcS and is in the process of isotopically labeling the proteins.

Aim 2. Tag central proteins of carboxysome and ABC transporter complexes (1/03). Molecular biology of producing tagged genes of the carboxysome is ongoing.

Aim 1. PCR amplify genes to be used as receptors in phage display. Design phage libraries. Begin testing 2/03). Genes have been cloned and libraries synthesized. Protocols are under development to begin selection.

Aim 3. Prepare antibodies. 10 genes. Characterize antibodies (3/03). We have synthesized and characterized four antibodies (two putative carbonic anhydrases, csoS2, and pepA).

Aim 3. Test improved accuracy of new scanner with labeling printed DNA with separate fluorophore (4/03). Complete.

Aim 2. MS characterize protein complexes. Determine consensus ligands (5/03). We have been testing protein content of soluble and particulate fractions using high-throughput LC-MS/MS and phosphate regulatory mutants.

Aim 3. Cross-calibration of microarrays. Submit gene expression data to ORNL group (8/03). We continue to collect RNA, hybridize, and analyze microarray gene expression data for phosphate regulatory mutants under conditions of phosphate depletion and various N- sources.

Aim 2. NMR sample conditioning and optimization for free proteins and protein-protein complexes with and without dilute liquid crystalline media (8/03). We have built and optimized our solution NMR facility.

Aim 3. Generate improved microarray data from statistically designed experiments (8/03). We continue to apply our successes with statistically designed yeast microarrays to improve the *Synechococcus* microarrays. We have begun to look at the full genome array data from TIGR to identify sources of error.

<u>FY04</u>

Aim 1. Finish phage display on other protein binding domains (1/04). Selection screens are ongoing.

Aim 2. Begin mutagenesis studies on proteins complexed in carboxysomes and ABC transporters (2/04). We have synthesized one cso operon knock-out mutant, knock-outs of putative carbonic anhydrase mutants, and four phosphate regulatory knock-out mutants.

Aim 3. Apply hyperspectral scanner to *Synechococcus* gene microarrays with multiply-tagged cDNA (7/04). This Aim is ongoing.

<u>FY05</u>

Aim 2. PCR amplify novel binding domains of carboxysome and ABC transporter proteins (10/04). We have completed research in PDZ domains as well as TPR domains and will begin cloning after initial phage-display experiments are complete.

Aim 2. Structural characterization of protein-protein complexes (4/05). We have successfully isolated carboxysomes and are beginning cryoelectron tomography studies to elucidate 3-D structures.

Aims 1, 2 & 3. Manuscript preparation (8/05). These Aims are ongoing. Also, see publications list.

Collaboration With Others

We are continuing our collaboration with Dean Price and Murray Badger of the Australian National University, and with Grant Jensen at Cal Tech to do electron microscopy and cryoelectron tomography studies of the carboxysome. This quarter we have started conversations with Dave Hanson at the University of New Mexico. Dr. Hanson recently completed a post-doctoral appointment in Murray Badger's lab and is interested in various metabolic processes in cyanobacteria.

We maintain strong interactions with GTL team members. We continue to work with Arie Shoshani, LBNL (Subproject 5) to develop a laboratory data-management platform. We have worked closely with the University of Michigan to develop experimental plans. Also, we are discussing future docking calculations with Steve Plimpton, SNL (Subproject 2), and his team.

There is continuing collaboration with the Werner-Washburne lab at University of New Mexico to test our hyperspectral microarray scanner with their yeast genome arrays. Collaborations established though Eugene Kolker (BiaTech) with other GTL laboratories (ORNL, PNL, Michigan State) to perform hyperspectral imaging and analysis for quality control are continuing. We have begun collaborating with Wim Vermaas's group at ASU to identify and map *Synechocystis* native pigments using our hyperspectral scanning technology.

We are collaborating with GTL "Center for Molecular and Cellular Systems," Frank Larimer and Michelle Buchanan, ORNL, to include the developed NMR-based technology into a planned set of biophysical characterization tools for DOE Facility III.

Publications and Presentations

Haaland, D. M., Timlin, J. A., Sinclair, M. B., Van Benthem, M. H., Keenan, M. R., Thomas, E. V., Martinez, M. J., Werner-Washburne, M., Palenik, B., Paulsen, I. "Improving Microarray Analysis with Hyperspectral Imaging, Experimental Design, and Multivariate Data Analysis." *2004 Genomes to Life Program Workshop*. Washington, D.C. February 29-March 2, 2004.

Haaland, D. M., Timlin, J. A., Melgaard, D. K., Keenan, M. R., Van Benthem, M. H., Sinclair, M. B. "Spectroscopy and Chemometrics: A Personal Vision for the Future." *Pittsburgh Conference, Bomem-Michelson Award Address*. Invited. Chicago, IL. March 7-12, 2004.

Haaland, D. M. "Multivariate Curve Resolution Applied to the Analysis of Hyperspectral Images." *Naval Research Laboratories*. Invited. Washington, D. C. May 19, 2004.

Timlin, J. A., Sinclair, M. B., Haaland, D. M., Martinez, M. J., Manginell, M., Brozik, S. M., Guzowski, J. F., Werner-Washburne, M. "Hyperspectral Imaging of Biological Targets: the Difference a High Resolution Spectral Dimension and Multivariate Analysis Can Make." *Conference Proceedings, IEEE International Symposium on Biomedical Imaging*. Arlington, VA. April 15-18, 2004.

Subproject 2: Computational Discovery and Functional Characterization of *Synechococcus* Sp. Molecular Machines

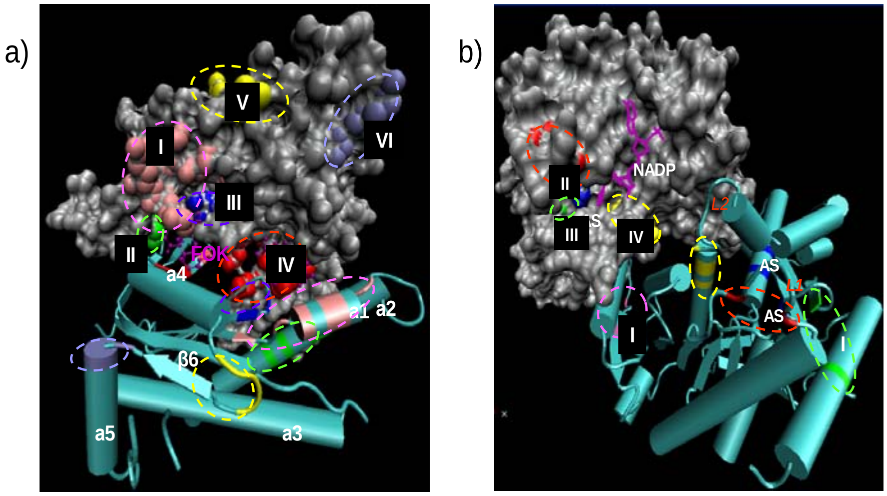


Figure 2-1. Mapping of key subpatches specific to guanylyl cyclases (GC) and adenylyl cyclases (AC) (a) and lactate dehydrogenases (LDH) and malate dehydrogenases (MDH) (b) on 3-D protein structures. Both structures have two subunits, in which one subunit is drawn in surface and the other in cartoon. The key subpatches, along with enzyme activators and cofactors, are mapped in different colors. Literature analysis indicated that these key subpatches are corresponding to residues in domain interface: a(II) or interface residues that directly interact with enzyme activators: a(I, III), substrate binding: a(IV) and b(II), and regulatory: b(IV) and active loops: b(II).

The work continued on all main directions of our Molecular Machines Subproject. New algorithms were developed and programs released. We have released the first version of PPM-Charger – our novel algorithm for parent-ion charge determination in tandem mass spectrometry. We began work on the PPM-based algorithm for *de novo* peptide identification. The ROBETTA server is analyzing very large sets of genes and currently is heavily oversubscribed, with the requests waiting over 20 days. We are also reporting a new computational procedure that predicts such clusters of specificity-determining residues among highly homologous functional protein groups. This capability will be an important part of our future pipeline for the functional annotation of the *Synechococcus* Sp. genome and elucidation of its protein-protein interaction network. In biophysical studies, our goals during this quarter were 1) to perform molecular-scale computations of RuBisCO, a key enzymatic protein in the *Synechococcus* carboxysome, involved in the carbon-fixation process; and 2) to model peptide/protein binding in collaboration with the phage-display experiments of Subproject 1.

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).

Accomplishments

Computational Methods for Mass Spectrometry: In the last quarterly report we presented a new idea leading to a robust algorithm for a classical (and unresolved) problem in tandem mass spectrometry: parent-ion charge determination. During this quarter, we considered several new directions for investigating the problem. Specifically, we suggested new discriminating parameters, which can also be used for parent-ion charge classification. As a result of this work, we currently have a library of such discriminators. This capability gives us a flexibility to develop efficient and versatile charge-determination methods, suitable for a large set of diverse tandem MS instruments and/or experiments. The first version of our software (PPM-Charger) was released and is now being used in ORNL's Center for Molecular and Cellular Systems. The development of PPM-Charger will be continued.

The PPM-Charger project was presented at the annual ASMS meeting (American Society of Mass Spectrometry) and generated a great number of requests for downloading from the research organizations involved in proteomics research across the country. Currently it is distributed only to Genomes:GTL participants, but we are considering the possibility of distributing PPM-Charger to other research organizations.

de novo Peptide Identification: This is another strategic direction we are going to develop using PPM technology. The potential importance of *de novo* identification methods for the GTL proteomics projects was discussed in our previous quarterly reports. Efficient identification of b- and y-ions by PPM creates interesting algorithmic opportunities for *de novo* methods. In a preliminary study, we have investigated a possible implementation of the method (PPM-Chain) consisting of four stages:

1) compute P_{by} probabilities and select subset of peaks with the P_{by} value above a certain threshold T (everything from 0.2 to 0.9 could be considered);
2) add complementary peaks, extending the initial set (y-ions therefore converted into corresponding b-ions);
3) construct all possible connecting chains; and
4) score them by length and by product of P_b values in the nodes of the connecting graph (P_b is the probability that the given peak is b-ion).

Two questions should be answered in regard to this approach. First, what length of tags this approach can provide in principle, given that we can predict only some fraction of the existing b- and y-ions in the spectra, and some of the ions are actually always missing (about ~80% typically can be found in charge 2 spectra). Second, how well the scoring function works – in other words, how often the correct target will be ranked first.

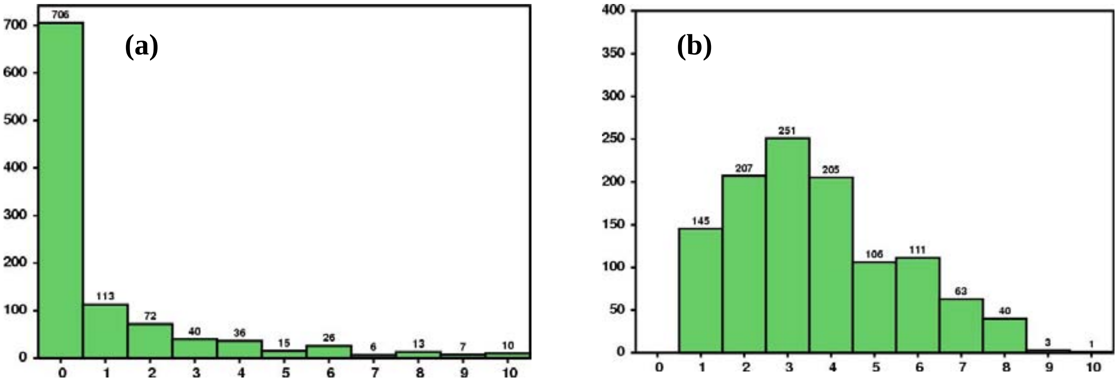


Figure 2-2. Statistics for the new algorithm of *de novo* peptide identification. (a) For 706 cases the "optimal tag" was at the top of the list (rank 0). (b) Length distribution in the obtained *de novo* tags.

Figure 2-2 contains some preliminary answers to these questions. Figure 2-2(a) presents a histogram of the ranks for "optimal tags" among ~1,500 spectra we used for this study. The "optimal tag" is the longest correct partial peptide possible for a given spectrum (this length is smaller than the full peptide length for a number of reasons, for example, because not all b- and y-ions are present in the given spectrum). The requirement for the tag to be an "optimal" one is stronger than the requirement to be a correct one, but as we can see in Figure 2-2(a), the optimal tag was at the top line of the list for 705 out of 1,534 tested cases. The existing algorithms for *de novo* peptide identification never reported a similar level of performance, as usually a very large list (>10-20 tags) had to be considered to provide a correct identification for ~50% of the cases.

In Figure 2-2(b), we have shown a histogram of length for correct peptide tags we obtained in the same data set. A significant number of those partial peptides equal or exceed the length 5. This is important because in bacterial genomes the tag of length 5 very often will be a unique one. Our capability to generate long peptide tags opens several attractive algorithmic opportunities, such as cross-validation between values of the parent mass, obtained sequence tag, and database. In the coming months we plan an extensive investigation and development of this novel algorithm, the release of version 1.0 of PPM-Chain for other Genomes:GTL projects, and the installation of the PPM-Chain web server.

Bioinformatics Methods for Prediction of Protein Interfaces: Protein interactions with other molecules play a central role in determining the functions of proteins in biological systems. Enzyme-substrate binding is a subset of these interactions that is of key importance in metabolic, signaling, and regulatory pathways. Bacterial chemotaxis, osmoregulation, carbon fixation, and nitrogen metabolism are just a few examples of the many complex processes dominated by enzyme-substrate recognition. One of the most remarkable properties of enzyme-substrate binding is high specificity. For example, the Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO; EC 4.1.1.39) is an enzyme that binds at a given site on a range of substrates with different compositions and shapes. RuBisCO catalyzes two separate reactions, carboxilation and oxygenation reactions depending on whether CO_2 or O_2 binds in its active site, respectively.

Identification of key residues or their clusters for substrate recognition (e.g., catalytic centers, substrate and cofactor binding sites, and hinge-motion controlling loops) is essential for elucidating the order and control principles of enzyme-substrate recognition. Their reliable prediction has immediate implications for drug design, protein engineering, elucidating molecular pathways through site-directed mutagenesis, and detailed functional annotation. X-ray and NMR structures of enzymes derived with their substrates can provide explicit specificity-determining residue information. Unfortunately, some macromolecular complexes do not easily yield x-ray-quality co-crystals and are often beyond the current limits of NMR spectroscopy. Site-directed mutational analysis remains a mainstay of specificity-determining site characterization. However, given the size of a typical enzyme (several hundreds of amino acids), it is difficult to know, *a priori*, the residues that are responsible for specificity. Locating these specificity-determining residues (SDRs) experimentally is still labor-intensive and time-consuming.

We developed a computational procedure that predicts such clusters of specificity-determining residues among highly homologous functional protein groups. Current methods identify conserved residues but largely ignore nonconserved residues and their potential contributions. Our method has the ability to overcome those limitations. For our case studies, we have investigated two highly homologous enzymatic protein pairs: guanylyl cyclases versus adenylyl cyclases and lactate dehydrogenase versus malate dehydrogenases, and applied this algorithm to plant and cyanobacterial RuBisCO protein complexes, which differ dramatically in the CO_2/O_2 specificity. Without using experimental data, we identified monoresidue clusters as well as multiresidue ones, and obtained a considerable concurrence with experimental results. Specifically, some of the identified clusters, primarily the monoresidue ones, can cover residues that are directly involved in substrate-enzyme interactions. Others, mainly multiresidue ones, cover residues vital for domain-domain and regulator-enzyme interactions, indicating potential roles of those function-nonspecific yet complementary residues in the specificity determination (Figure 2-1).

ROBETTA Server: During this quarter the ROBETTA server analyzed many more genes. The server was ported to another cluster, conjo.lanl.gov, and porting issues were ironed out. The ROBETTA server has become wildly popular and is in fact over-subscribed on-line. Current wait times are backed up 20 days and despite this users continue to submit sequences for prediction. Due to this greater than hoped for success, some funds may be reallocated to upgrading the disk storage and throughput. The successful port means that we may be able to replicate the server elsewhere. This is nontrivial and requires a great deal of “insider” knowledge, as it was not originally built with the portability requirement in mind. We are acquiring experience now that will lead to more effortless portability of this complex system in the future.

Our collaborators at the University of Washington have now ported the Rosetta code to C++ and all work on the FORTRAN branch is being merged. This will become the new system and permit greater flexibility in exploring memory management. This is particularly crucial to one of the intentions of this Subproject, in which ORNL and LANL plan to examine the feasibility of replica-Monte Carlo as an annealing/search method to replace the current metropolis algorithm. Without advanced memory management, maintaining replicas in the global name space would have been prohibitively difficult.

Peptide/Protein Docking Methodology: This quarter we focused on the combinatorial problem of solving for the best peptide sequence (using an energy metric). Each of the 5-20 peptide positions can have 20 possible amino acids. We use a mixed-domain representation with globally fixed backbone and sidechain rotamers (200 per amino acid, on average), which are allowed local (but not global) flexibility. To solve this problem, we developed a two-step process. The first step is to solve the problem of the best sequence given an approximation of the Boltzmann-weighted energies of each sidechain over each of its rotamer states. The second step finds the lowest energy structure for each lowest energy sequence by solving for the lowest energy rotamers for each sequence.

We added a feature to PDock to calculate a pseudo-Boltzmann-weighted average of the energy over all amino acid rotamers. Since the intermolecular term (peptide sidechain/protein) dominates the interaction energy for most amino acids, we ignore the intramolecular term (peptide sidechain-peptide-sidechain) for estimating these probabilities. We search for the amino acid sequence that minimizes these energies using a combinatorial branch-and-bound algorithm. This solver integrates ideas from protein side-chain structure prediction and is able (on test problems) to find optimal solutions within seconds. Furthermore, we have augmented this solver to enable the enumeration of all near-optimal solutions, which typically takes a few minutes to an hour (depending on the scope of the enumeration). We are currently working on extending this capability to use these (near-) optimal peptide sequences to predict the optimal set of rotamers for each peptide sequence. Mathematically, this is exactly the same problem, so we expect it can be solved quickly.

Molecular Dynamics Simulations of RuBisCO: We have performed long time-scale molecular dynamics (MD) simulations of a RuBisCO complex that includes the binding pocket for CO_2 and the 3-stage trapdoor mechanism for gating entrance and exit of small molecules to and from the pocket. In a simulation with explicit solvent and a fully flexible protein (~50,000 atoms), a CO_2 molecule in the closed binding pocket escaped through a “back-door” mechanism within a nanosecond of simulation time. This invalidates part of the Schlitter/Wildner theory that assumed gaseous ligands could not escape the pocket with the trapdoor closed. In a simulation with implicit solvent, we left a portion of the trapdoor gating proteins flexible and the remainder of the protein rigid. Combined with a replica-exchange MD option in our code, this enabled us to simulate many hundreds of nanoseconds of time to look for conformational changes in the gate proteins. We observed a partial closing event of the outermost portion of the gate, with the C-terminus initially coiling into an alpha-helix and subsequently stretching to clamp the trapdoor shut (see Figure 2-3). We are currently enhancing our MD code to enable even longer time-scale simulations and plan to compute binding pocket response to the presence of reactant and charged product molecules.

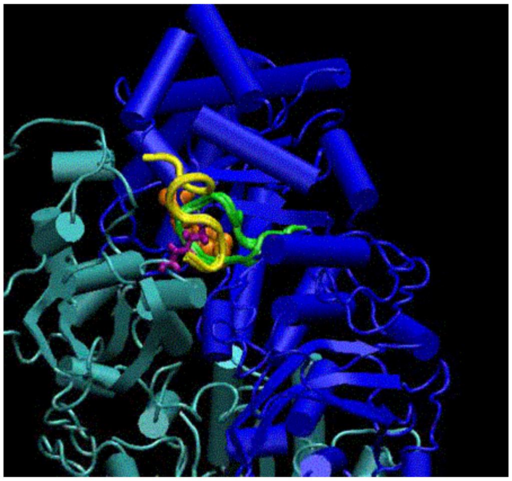


Figure 2-3: A portion of the RuBisCO protein complex (blue, gray) showing the binding pocket with a bound ligand (orange), and 3-stage trapdoor (purple, green, yellow). This snapshot is at the end of the simulation where the yellow C-terminus has formed into a helix and stretched across the pocket to a closed state.

Quantum Computations of RuBisCO Binding Pocket Effects: RuBisCO, like many proteins, prefers a particular ion (Mg) as an essential coenzyme. In an effort to understand why, we computed the hydration free energies for several ions, including Mg, Ca, Mn, and Zn. A surprising result was that in all cases the structures of the nearest shell of coordinated waters were the same – octahedral. We also found a partial explanation for RuBisCO's preference for Mg in terms of thermodynamic stability. It is less stable than many other divalent ion candidates, making it more likely to leave bulk water and bind in the protein's active site. Next, we plan to examine the stabilities and structures of Mg and other ions in the active site of RuBisCO.

Proton transfer initiates activity in RuBisCO. In order to understand enzyme activity, it is important to predict the dynamic changes in protonation states that arise as reactions occur. We examined proton transfer in glycine, a model for the catalytic CO_2 reaction occurring in RuBisCO, using an *ab initio* approach (no fitting of parameters), and validated our studies against experimental data. We successfully predicted the most-stable protonation state, and established a rate for one of the proton transfer reactions. Next, we plan to examine components of the RuBisCO active site and make similar predictions.

Progress Towards Milestones

Aims 1, 2. Implement incorporation of the experimental restraints (NMR and mass spectroscopy) in all modeling tools; explore various regimes of experimental data integration and application (4/04). We are continuing development of tools for mass spectrometry. This quarter, we report preliminary studies for the novel *de novo* peptide identification algorithm.

Aim 2. Continue simulation of ligand library conformations for phage display ligands and appropriate mutants (4/04). We have developed several new capabilities with direct impact to the combinatorial problem of finding the best peptide ligand, including an implementation of pseudo-Boltzmann-weighted average of the energy over all amino acid rotamers and a new solver enabling the enumeration of all near-optimal solutions.

Aims 3, 4. Develop tools for constructing protein-protein interaction maps (4/04). A new algorithm for identification of key residues involved into substrate recognition will be very helpful for elucidating the order and control principles of enzyme-substrate reactions.

Collaboration With Others

In the development of mass spectrometry analysis capabilities, we continued our collaborative effort with ORNL-PNNL Genomes to Life project, Center for Molecular and Cellular Systems. Our projects involved Edward Uberbacher, Frank Larimer, Gregory Hurst, Robert Hettich, Hayes McDonald, David Tabb, Brad Strader, Tema Fridman, and Jane Razumovskaya. We had constructive discussions and are planning joint projects with Gordon Anderson and William Cannon from PNNL. New collaborative effort is underway with George Church's Genomes:GTL project. This effort will be centered on development of new computational capabilities for proteomics.

Computational molecular biophysics collaborations included several discussions with Professors Guenter Wildner and Juergen Schlitter (University of Bochum, Germany) who are experts in RuBisCO protein function and dynamics. Prof. Wildner visited Sandia as part of this collaboration. We are working with James Critchley at RPI on rigid-body dynamics algorithms appropriate for our MD code. ROBETTA server development has been done in collaboration with David Kim and Dylan Chivian, members of the David Baker lab at the University of Washington.

Publications and Presentations

Asthagiri, D., Pratt, L. R., Paulaitis, M. E., Rempe, S. B. "Hydration Structure and Free Energy of Biomolecularly Specific Aqueous Dications, Including Zn2+ and First Transition Row Metals." *JACS*, 126, 1285-1289. 2004.

Gorin, A., Day, R. M., Razumovskaya, J., Fridman, T. "Probability Profile Method – Novel Approach for Construction De Novo Tags in MS/MS Experiments." *Keystone Conference in Mass Spectrometry*. 2004.

Leung, K., Rempe, S. B. "Ab Initio Molecular Dynamics Study of Formate Ion Hydration." *JACS*, 126, 344-351. 2004.

Leung, K., Rempe, S. B. "Ab Initio Molecular Dynamics Study of Glycine Tautomerization in Water." Submitted. 2004.

Razumovskaya, J., Fridman, T., Day, R., Uberbacher, E., VerBerkmoes, N., Hettich, R., Borziak, A., Gorin, A. "Charge Determination in Low-Resolution Tandem Mass Spectrometry." *Proceedings of American Society Mass Spectrometry*, Accepted. 2004.

Rempe, S. B. "Ab Initio Molecular Dynamics of Glycine Tautomerization in Water." *Biophysical Society Meeting*. Baltimore, MD. February 2004.

Rempe, S. B. "Ab Initio Molecular Dynamics of Glycine Tautomerization in Water." *American Chemical Society Meeting*. Anaheim, CA. March 2004.

Rempe, S. B., Asthagiri, D., Pratt, L. R. "Inner Shell Definition and Absolute Hydration Free Energy of K+(aq) on the Basis of Quasi-Chemical Theory and Ab Initio Molecular Dynamics." *Physical Chemistry Chemical Physic*s, 6, 1966-1969. 2004.

Yu, G-X., Park, B-H., Chandramohan, P., Munavalli, R., Geist, A., Samatova, N. F. "In-silico Prediction of Surface Residue Clusters for Enzyme-Substrate Specificity." *IEEE Bioinformatics Conference*, Submitted. 2004.

Subproject 3: Computational Methods Towards The Genome-Scale Characterization of *Synechococcus* Sp. Regulatory Pathways

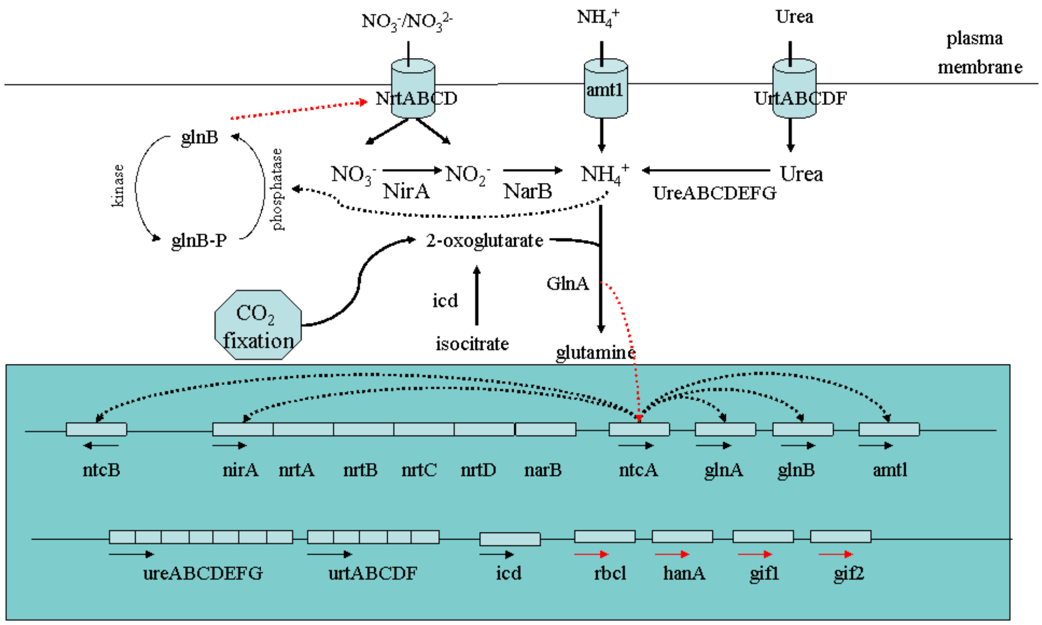


Figure 3-1. A predicted regulatory network (part) of nitrogen assimilation of *Synechococcus* Sp. WH8102.

Progress continues in our effort to develop new and effective protocols for tackling the challenge of systematic characterization of regulatory pathways. We continue to develop and refine our computational pipeline for microbial pathway inference and reconstruction. In particular, we continue to refine and improve the following techniques required for regulatory pathway/network inferences: a) biological pathway/network mapping across different genomes; b) prediction of regulatory binding sites; c) prediction of terminators of operons; d) prediction of operons; e) prediction of regulons; f) microarray data analysis and mining; g) mining of biological literature; and h) applications of our computational prediction capabilities to more pathway/network inferences in *Synechococcus* Sp. WH8102.

This work was funded in part or in full by the US 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).

Accomplishments

We are making rapid progress towards our goals of systematic inference of signaling and regulatory pathways via data mining of high-throughput biological data of various kinds coupled with computational modeling and experimental validation. We are in the process of completing a number of method developments in protein-protein interaction predictions, operon structure prediction, transcription-factor binding-site identification, mapping of known biological pathways, statistical analysis of microarray gene expression data, and automated literature mining.

Development of a Computational Capability for Mapping Known Biological Pathways from One Genome to Another Genome: One key piece of information we are using to derive biological pathways and networks is gathered through mapping known (partial) homologous pathways and networks from one genome to our target genome. We have implemented two algorithms to accomplish this mapping. Both algorithms employ homology information and genomic structure information, including operon and regulon structures in the mapping, a key distinguishing feature of our algorithms from other existing algorithms. We have demonstrated that using such combined information makes pathway/network mapping much more accurate than using homology information alone, a strategy typically employed by existing algorithms. The two algorithms differ in their detailed computational formulation and solution. In the first algorithm, the problem is formulated and solved as a Steiner Network problem, while in the second and improved algorithm, the problem is formulated and solved as a Integer Programming problem. The improved algorithm allows us to map a large number of pathways and networks (at genome scale) simultaneously from one genome to another. Both algorithms have been implemented as part of the P-MAP software. Using this capability, we are in the process of mapping many known E. coli pathway and networks to *Synechococcus* Sp. WH8102.

Development and Improvement of a New Capability for Prediction of Regulatory Binding Sites at Genome Scale: Through comparative genome analyses, we have extracted the upstream regions of the orthologous genes across all sequenced cyanobacteria genomes. We then predicted the conserved sequence motifs in these regions, using our own program CUBIC and Bio-Prospector. We clustered the predicted binding sites to filter out the accidental predictions of erroneous motifs. The clustering problem of predicted binding sites proves to be a highly challenging problem, due to a number of factors, including a) overlaps of predicted binding sites with different sizes; and b) assessment of statistical significance of clustering results. A number of algorithms are currently being explored to make the predictions of binding sites at genome scale, a very challenging problem, more reliable.

Prediction of Operon Structures: We have developed a comparative genomics approach for predicting operons in *Synechococcus* Sp. WH8102 that combines many known characteristics of an operon structure concerning the functions, intergenic distances, and transcriptional directions of genes, promoters, terminators, etc. in a unified likelihood framework (Chen, et al., 2004a). The data and results are available to the public at <http://www.cs.ucr.edu/~xinchen/operons.htm>. Our work in the last quarter was focused on improving the approach from several aspects. First, we have removed the intergenic distance threshold requirement and incorporated the distance information into the likelihood formula. Second, instead of considering the conservedness of genes in a small number of genomes, we are now using phylogenetic profiles of genes in a large number of bacterial genomes. Third, we are trying to refine the functional category information by using the GO database, although this has been challenging due to technical issues.

Prediction of Orthologous Genes: The identification of orthologous genes is a fundamental problem in comparative genomics and evolution and is very challenging, especially on a genome-scale. We have been working on a new approach for assigning orthologs between different (but related) genomes based on homology search and genome rearrangement. The preliminary experimental results on simulated and real data demonstrate that the approach is very promising (it is competitive to the existing methods based only on homology), although more experimental validations need to be done (Chen, et al., 2004b). At the moment, our algorithm only works for single-chromosomal genomes, but we are working on an extension to multi-chromosomal genomes and consideration of more realistic biological constraints.

Prediction of Operon Terminators: We have recently implemented a software package, Rnall, for Rho-independent terminator (RIT) prediction. Rnall is able to scan a microbial genome sequence for RNA secondary-structure prediction. To achieve high computational efficiency, Rnall scans along the nucleotide sequence with a sliding window and extracts possible local secondary-structure candidates based on dynamic programming of a simple scoring scheme. Then Rnall uses thermodynamic energy parameters to optimize local secondary structures. The computational time complexity of Rnall is $O(LW^2)$ with a space requirement $O(L+W^2)$, where L is the length of nucleotide sequence, and W is the sliding window size. The software Rnall is available upon request.

We applied Rnall to screen RITs in E. coli and all the 101 non-redundant RITs recorded in RegulonDB were successfully retrieved. We are comparing the results of Rnall with results by several other RIT prediction software packages, including TransTerm, GesTer, and RNAMotif. We carried out a systematic study of RITs for *Synechococcus* Sp. based on Rnall in conjunction with comparative genomic studies using seven other related cyanobacteria whose complete genomic sequences are available. From the 101 protein-coding genes with known RITs in E. coli, we were able to detect 40 E. coli homologous genes in cyanobacteria. Multiple sequence alignment and phylogenetic footprint of the RITs in different genomes reveal interesting evolutionary patterns in conservation and divergence. In particular, although some conservation patterns of the RITs among orthologous genes are observed, a gene whose ortholog has an RIT may not have an RIT for itself. We have built a model to predict the confidence of RIT based on the pattern of several known RITs across the genome, include pflB and araD. We have surveyed the RITs using Rnall, in *Synechococcus* Sp. and the seven other related cyanobacteria, and a cluster of RITs with different confidence levels are predicted.

Development of a Local Database for Prediction Results of Synechococcus Genome: We have developed a *Synechococcus* WH8102 knowledge base (http://www.csbl.bmb.uga.edu/WH8102), which is a web-based relational database developed to facilitate computational effort to reconstruct regulatory pathways and serve as a gateway for biologists to access the data. It is the repertoire that integrates a variety of knowledge derived both from literature and computational prediction. Those data are organized in hierarchical fashion. The basic building blocks are functional annotation and structure prediction of individual molecules. Those data are then organized into clusters based on computationally predicted operon, regulon, and molecular complexes. Finally, all data are complied into pathways derived from combined efforts of literature mining and computational prediction. A number of tools have been developed to facilitate the data retrieval, including a SQL query engineer and several viewers to browse genomes, molecular complexes, and pathways.

Microarray Chip Design and Data Analysis: A number of whole-genome *Synechococcus* microarrays were printed and hybridized at TIGR. In one set of microarrays, each gene was replicated six times in close proximity on the array. For this situation, a statistical analysis was conducted to assess the degree to which conclusions regarding differential expression of genes (experimental versus control) might be affected by the “technical variability” observed both within and across arrays. As expected for this situation, the intensity-ratios of spots within an array were quite repeatable. Slide-to-slide repeatability for this situation was inconsistent. Therefore, in some cases, conclusions regarding differential expression of genes would likely vary depending on the slide. Because the original whole-genome print design crafted by Sandia was not realized in the first print, another set of *Synechococcus* microarrays were printed and hybridized at TIGR where the six gene replicates were spread out widely across the array. We intend to compare the “within-array” level of intensity-ratio variation for this set of arrays with the level of within-array and across-array variation previously observed. Initial experiments to use the new microarrays have been planned and will be carried out in the next quarter.

We have developed and implemented in MATLAB® a significantly faster maximum-likelihood principal component analysis (MLPCA) algorithm applicable to a data matrix containing more variables than samples and where the samples comprising the data matrix share an equal row-error covariance matrix. For a data matrix containing more variables than samples, our algorithm operates on the smaller sample-by-sample dimensioned-error covariance matrix, as opposed to the variable-by-variable dimensioned-error covariance matrix as advocated in current MLPCA algorithms. Our modified MLPCA algorithm for equal row-error covariance can provide orders of magnitude improvement in speed for analyzing microarray-based gene expression datasets where the number of genes or variables significantly exceeds the number of patients or samples. Our algorithm allows the processing of datasets containing many variables (greater than 7,000), which produce out-of-memory errors employing current MLPCA algorithms.

Gene selection tests were completed on Affymetrix human leukemia microarray data while we await adequate numbers of whole-genome *Synechococcus* microarrays to become available. The t-test, Wilcoxon rank sum, and jack-knifing methods were compared for their ability to select genes to predict remission versus failure of chemotherapy treatment using an archival leukemia microarray dataset composed of 164 training samples (71 remission/93 failure), 83 test samples (38 remission/45 failure) and 12,581 genes. The best three and best seventeen genes for differentiating the remission and failure observations in the training set were selected using the t-test, Wilcoxon rank sum, and jack-knifing methods. Classification success rates of partial least squares remission/failure classification models using all 12,581 genes and selected sets of three and seventeen genes were estimated using leave-one-out cross-validation (CV) of the training samples. Using all 12,581 genes, a baseline, CV-based classification success rate of 62.2% was obtained for the training samples. When the best three genes selected by the t-test, Wilcoxon rank sum, and jack-knifing methods were used in developing the classification model, similar CV-based classification success rates of 74.39%, 69.51%, and 71.34% were obtained for the training samples. In using the best set of seventeen genes, however, the jack-knifing method (classification success rate of 89.6% for the training samples) was superior to the t-test (78.1% classification success rate) and Wilcoxon rank sum (79.3%) methods. This result suggests that a multivariate feature-selection method is preferable to a univariate feature-selection method for microarray datasets where remission and failure groups can be distinguished based only on subtle, correlated differences in gene expression levels, which is characteristic of the leukemia dataset.

Informatics for Literature Mining Support: The validation of proposed networks from automatic inference methods includes the very time consuming step of reading and evaluating the literature with respect to the proposed connections. We have begun addressing ways to make this time- and knowledge-intensive task somewhat easier. We are interested in saving time and effort by automatically weeding out papers that appear to be relevant at initial glance, but which ultimately prove to have no useful connection to the genes, processes, and mechanisms associated with the nodes in the inferred network. We have established a contract with the computational linguists at the Computer Science Laboratory at New Mexico State University to help us build the required tools.

We begin by taking the results from typical keyword-based queries to the ISI database of scientific papers. These papers are the seeds around which we find dozens of similar papers within an XY ordination of publications derived from co-citation similarities (co-citation methods have certain strengths over keyword based groupings). These sets are then examined with computational linguistics tools to identify strong themes across the papers in the group. Often, we find that the original keyword query only incidentally relates to strongest themes in the set of abstracts. We have two direct uses for these strong themes. First, they are useful in deciding which groups of papers are truly relevant to our original queries. Second, we believe that they can be used to create good summaries of small collections of potentially relevant abstracts.

To date, we have assembled all of the abstracts indexed by ISI in 2002 and have clustered them with the VxOrd tool. This cluster is one of the largest ever completed and includes 833,000 papers and 5.6 million similarity links. We have identified the 20 abstracts in the neighborhood around each of 118 papers matching a query for “carbon fixation.” The abstracts in these groups were examined manually and overall summaries for each cluster were created from these abstracts, again manually. We compared these manually generated summaries to those created automatically using the Microsoft Word summarization tool. The automatically generated summaries were deemed to be inadequate relative to the manually created ones and relative to our actual needs. We are continuing to explore the more experimental tools and different methods that our collaborators at NMSU believe may be able to meet our needs.

Progress Towards Milestones

Aim 1. Improved technologies for information extraction from microarray data (10/04). We have tested the within- and between-microarray repeatability of whole-genome *Synechococcus* microarray data. While the within-array repeatability was good, some inconsistencies were found with the between-array repeatability.

Aim 2. Improved capabilities for analysis of microarray gene expression data (9/04). We have dramatically speeded up the performance of the maximum-likelihood principal component analysis (MLPCA) algorithm that can be used for analysis of microarray data. Two univariate methods and one multivariate method of gene selection were tested with experimental microarray data, and the multivariate jack-knifing method clearly out-performed the univariate methods for selecting genes to predict remission versus failure of chemotherapy treatment.

Aim 3. Improved capabilities for binding-site identification and application of this capability for binding-site prediction (7/04). We have developed a novel computer program, CUBIC, based on our minimum-spanning, tree-based, data-clustering framework for identification of regulatory binding sites. In the past quarter, we continued to refine and extend this capability to genome-scale prediction of regulatory binding sites. We have applied this capability to predict the regulatory binding sites of transcription

Aim 5. Improved capability for inference of biological pathways (10/04). We have developed and continue to improve a system for computational prediction of biological pathways. The system consists of four major components: a) pathway template construction in genomes with a great amount of data available; b) pathway mapping through orthologous gene mapping under constraints of preserved operon and regulon structures; c) pathway model refinement and expansion; and d) final pathway model prediction and suggestion for experimental validation. We have made rapid progress in building up strong prediction and modeling capabilities in each of the four areas. Using this system, we have predicted phosphorus assimilation pathways, carbon fixation and assimilation pathways, and nitrogen assimilation pathways. We expect many more *Synechococcus* WH8102 pathways will be predicted using this capability.

Aim 6. Refine approaches for scanning and analyzing our DNA microarrays; provide slides that we have scanned for interlab calibration (10/03). In another phase of our investigation, fluorescence data from replicate cDNA microarrays from the KUGR Microarray Facility (UNM) were analyzed to further understand the limitations and problems associated with microarray data obtained from printed cDNA arrays. A number of graphical methods for detecting processing anomalies were developed and used to explore the data. We found that background levels were relatively high and spatially variable, suggesting inadequate blocking of the amine coating outside the printed spot. Some problems associated with cDNA printing were also discovered. In addition, there was sporadic block-dependent slide-to-slide dye variation. We have found that replicate control spots (printed in every array block) are useful for detecting these anomalies.

Aim 7. Capture knowledge from our biological collaborators, in close collaboration with the computational linguists, to develop programs to read and begin to understand the relevant text (08/03). Preliminary capabilities are being developed to extract information from the published abstracts to assist scientists in rapidly screening a large pool of scientific abstracts for relevant research articles.

Collaboration With Others

Our Subproject 3 team has been in close contact with Brian Palenik of UCSD/Scripps Research Institute (Subproject 1) and Dong Xu, University of Missouri, as well as Andrey Gorin, ORNL (Subproject 2), Nagiza Samatova, ORNL (Subproject 2), Al Geist of ORNL (Subproject 5), and Frank Olken, LBNL (Subproject 5).

Publications and Presentations

Chen, X., Su, Z., Dam, P., Palenik, B., Xu, Y., Jiang, T. “Operon Prediction by Comparative Genomics: an Application to the *Synechococcus* sp. WH8102 genome.” *Nuclear. Acids Res*. 32 (7), 2147 – 2157. 2004.

Chen, X., Zheng, J., Fu, Z., Nan, P., Zhong, Y., Lonardi, S., Jiang, T. “Assignment of Orthologous Genes via Genome Rearrangement.” Submitted. 2004.

Dam, P., Su, Z., Chen, X., Olman, V., Jiang, J., Xu, Y. “*In silico* Construction of the Carbon Fixation Pathways in *Synechococcus sp.* WH8102.” *Proceedings of the Third IEEE Computer Science Society Conference on Bioinformatics (CSB04).* In press. 2004.

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

Haaland, D. M., Timlin, J. A., Melgaard, D. K., Keenan, M. R., Van Benthem, M. H., Sinclair, M. B. “Spectroscopy and Chemometrics: A Personal Vision for the Future.” *Pittsburgh Conference, Bomem-Michelson Award Address*. Invited. Chicago, IL. March 7-12, 2004.

Haaland, D. M. “Multivariate Curve Resolution Applied to the Analysis of Hyperspectral Images.” *Naval Research Laboratories*. Invited. Washington, D. C. May 19, 2004.

Mao, F., Su, Z., Olman, V., Chung, D., Xu, Y. “Pathway Mapping With Operon Information: An Integer-Programming Method.” *Proceedings of the Third IEEE Computer Science Society Conference on Bioinformatics (CSB04).* In press. 2004.

Olman, V., Peng, H., Su, Z., Xu, Y. “Mapping of microbial pathways through constrained mapping of orthologous.” *Proceedings of the Third IEEE Computer Science Society Conference on Bioinformatics (CSB04)*. In press. 2004.

Olman, V., Peng, H., Su, Z., Xu, Y. “Mapping of microbial pathways and regulons across multiple genomes.” Submitted. 2004.

Su, Z., Dam, P., Chen, X., Olman, V., Jiang, T., Xu, Y. “Computational construction of nitrogen assimilation pathway in cyanobacteria *Synechococcus* sp. WH8102.” *Proceedings of the Third IEEE Computer Science Society Conference on Bioinformatics (CSB04)*. In press. 2004.

Su, Z., Dam, A., Chen, X., Jiang, T., Palenik, B., Xu, Y. “Computational Inference of Regulatory Pathways in Microbes in *Synechococcus sp.” Genome Research*. Under review. 2004.

Wang, P., Su, Z., Xu, Y. “A knowledge base for computational pathway reconstruction in *Synechococcus sp* WH8102.” *Proceedings of the Third IEEE Computer Science Society Conference on Bioinformatics (CSB04).* In press. 2004.

Xu, Y. “Computational challenges in inference of microbial regulatory networks.” Biology Department, Georgia Tech. Invited by Prof. Mark Borodovsky. April 12, 2004.

Xu, Y. “Computational Methods for Inference of Microbial Regulatory Networks.” Seminar. Academia Sinica, Taiwan. May 18, 2004.

Xu, Y. “Computational Challenges in inference and reconstruction of microbial regulatory pathways.” “IEEE Conference on Bioinformatics and Bioengineering (BIBI’04).” Invited. Taichung, Taiwan. May 20, 2004.

Xu, Y. “A Computational Methods for Protein Mass Spectrometry Data Interpretation.” Tsinghua Bioinformatics Workshop. China. June 7, 2004.

Subproject 4: Systems Biology for *Synechococcus* Sp.

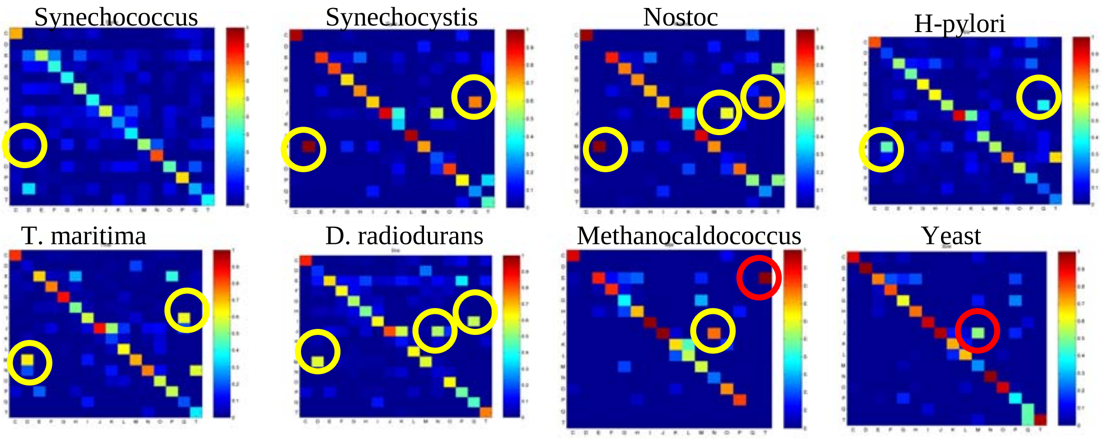


Figure 4-1. Functional linkages. The plots depict the number of linkages found between protein functions in eight organisms. The protein functions are those of the COG database, and comprise three information storage and processing functions (J, K, L), six cellular processes functions (D, O, M, N, P, T), and seven metabolism functions (C, G, E, F, H, I, Q). Yellow circles depict functional linkages found across different organisms, and red circles are linkages unique to an organism.

Within our top-down approach (Aim 1), we have used our inference and analysis network tools to functionally characterize eight organisms, including six bacteria (Figure 4-1). We find that functional distances between the studied organisms match their phylogenetic distances well. We also observe that there are binding rules between protein functions, that is, most proteins bind to proteins having the same function, while interfunction bindings occur rarely. Related to our bottom-up approach, we have compared the precision versus sensitivity of our signature-product method with the technique recently published by Jansen, et al. (2003). The results obtained using the signature product are an improvement over those published by Jansen, et al. We have also demonstrated parallelization of our discrete component model, ChemCell, at an efficiency of >60% on 512 processors. For larger systems, it is anticipated that even greater efficiency will be reached. Finally, we have continued to refine our linkage between hierarchical levels and are now testing component levels against the data, such as that available from the HOT (Hawaiian Ocean Time-series) program. We continue to broaden our incorporation of domain experts under our *encapsulation of expertise* paradigm by now including the work of Grover (2003), Geider, et al. (1998), and Litchman and Klausmeier (2001) for modeling the mapping between extracellular ocean nutrient concentrations (carbon, nitrogen, and phosphorus) and intracellular cell quotas, as well as temporal oscillations in irradiance. We are working to leverage our work to the benefit of other programs by contributing as co-Principal Investigators under the leadership of Dr. Mark Boslough of Sandia on an LDRD proposal entitled “Coupling Biogeochemistry into Climate Models with Macroecological Scaling.”

This work was funded in part or in full by the US 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).

Accomplishments

Protein Interaction Network Inference and Analysis: In the top-down approach, we have used the inference and analysis tools developed last quarter to perform a functional analysis of eight different organisms (six bacteria, one archea, and one eukaryota). To be precise, using the COG database (http://www.ncbi.nlm.nih.gov/COG), we functionally classified the proteins involved in *Synechococcus* WH 8102, *Synechocystis* PCC 6803, *Nostoc* PCC 7120, *T. maritima* 2336, *D. radiodurans* 1299, *H-pylori* 26695, *Methanocaldococcus* 2190, and *Yeast* 4932. The *Synechococcus* network used was the one inferred last quarter within Subproject 3, and all other networks were extracted from the STRING database (http://string.embl.de). We find the functional composition differences of these organisms to be consistent with their phylogenetic distances. For instance, the smallest functional distance for *Synechococcus* is *Synechocystis*, which is also the closest organism phylogenetically. Using the protein interaction networks, we compiled the functional linkages for the eight studied organisms. For all organisms, we find that proteins tend to bind to proteins within the same functional category (cf. intrafunctional linkages in diagonals in Figure 4-1), while most interfunctional linkages do not occur (cf. function C in Figure 4-1, which binds only to itself). We also observe that proteins involved in translation and biogenesis are the strongest binders (cf. function J in Figure 4-1). Finally, we notice that some interfunctional linkages in bacteria are not found in archea and eukaryota, while conversely, archea and eukaryota have unique interfunctional linkages. Figure 4-1 seems to indicate that there are some binding rules between protein functions. We plan to use these rules to validate the networks inferred with our bottom-up approach.

For the bottom-up approach this quarter, we compared our signature-product technique with the method in Jansen, et al. (2003). Both our method and Jansen's method are supervised in the sense that examples of protein-protein interactions must be provided in order for either method to make predictions. Our methods differ in that the variables we use are based on sequence information, while the variables used in Jansen are based on combinations of other sources of information (such as experiment and de novo calculations). The methods are quite different and so a comparison is interesting.

To make a comparison between our method and Jansen's method, we downloaded the Yeast predictions made by Jansen from their website and computed the sensitivity and precision using different thresholds of confidence. These thresholds give the confidence of the method in its predictions, where a higher threshold indicates a higher confidence. By varying the threshold we can be more certain of making correct positive predictions (interactions), but less certain of picking up every positive prediction. This trade-off is captured by comparing the precision and sensitivity statistics, where (roughly) the precision gives the accuracy on positive predictions, and the sensitivity gives the percent of positives detected.

Since our method also includes a confidence threshold, we are able to compare the precision/sensitivity trade-off of our method with that of Jansen. This comparison is shown in Figure 4-2, where the PIE, PIP, and PIT curves refer to different predictions made in Jansen, et al. (2003).

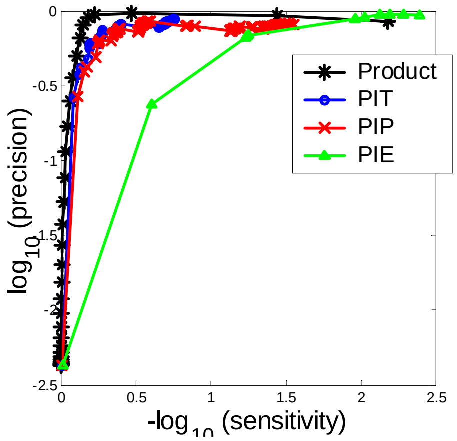


Figure 4-2. Precision versus sensitivity for Yeast protein interaction network. The Jansen, et al. PIE predictions are based on experimental input, the PIP predictions are based on input from de novo computations, and the PIT predictions are based on both experimental and de novo input. Product refers to our signature-product technique.

Discrete Component Model: This quarter we have continued development and testing of our particle-based ChemCell model of protein/protein interactions within cells. In ChemCell, individual particles represent protein molecules or protein complexes. They diffuse through the simulation geometry via Brownian motion and pair with nearby particles using Monte Carlo rules to react according to an input list of possible chemical-reaction equations.

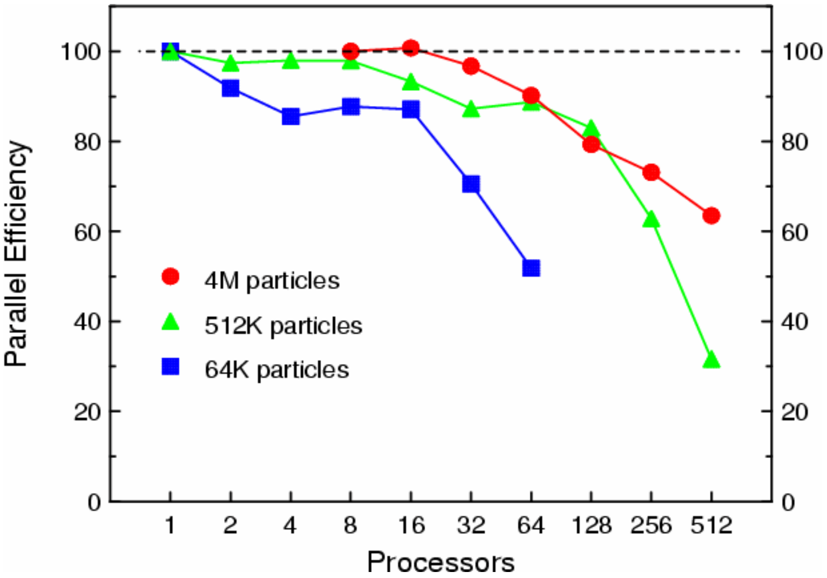


Figure 4-3. Parallel-scaling efficiency of ChemCell for different sized systems.

We tested the parallel-execution and load-balancing capabilities of ChemCell on a series of model problems ranging from a few thousands of particles up to several million (see Figure 4-3). Reasonable scalabilities of over 50% parallel efficiency were obtained even for fixed-sized problems on many processors. For example, a 4-million particle problem was 60% efficient on 512 processors of Sandia’s Intel Tflops machine (about 300-times faster than the same run on a single processor).

Computational performance of the code was tested and runs of a simplified network of glucose production in *Synechococcus* were made for use in a variety of presentations in the last quarter, as listed below.

<u>*Continuum System Models for Synechococcus:*</u> We are now pursuing data on experimentally observed aggregations of *Synechococcus* due to concentrations of nutrients. We intend to establish the relative mean distances between these species in their environment; the length scales of interest here are crucial to assessing the impact of multiple *Synechococcus* uptaking and fixing CO_2. Concurrently, we have embarked upon efforts to apply the already modified capabilities of MPSalsa to the larger-scale ecological models under development in the hierarchical modeling portion of the project. The same convective transport of nutrients and CO_2 already applied to the aggregations of *Synechococcus* on an individual cellular level are of significant interest at the population scale (Figure 4-4). Over sections of ocean on scales of meters, as opposed to microns, we are now implementing the developments of the hierarchical modeling population models in MPSalsa, whose spatial simulation capabilities allow exploration of these larger-scaled transport mechanisms on growth patterns and nutrient consumption as well as CO_2 fixation.

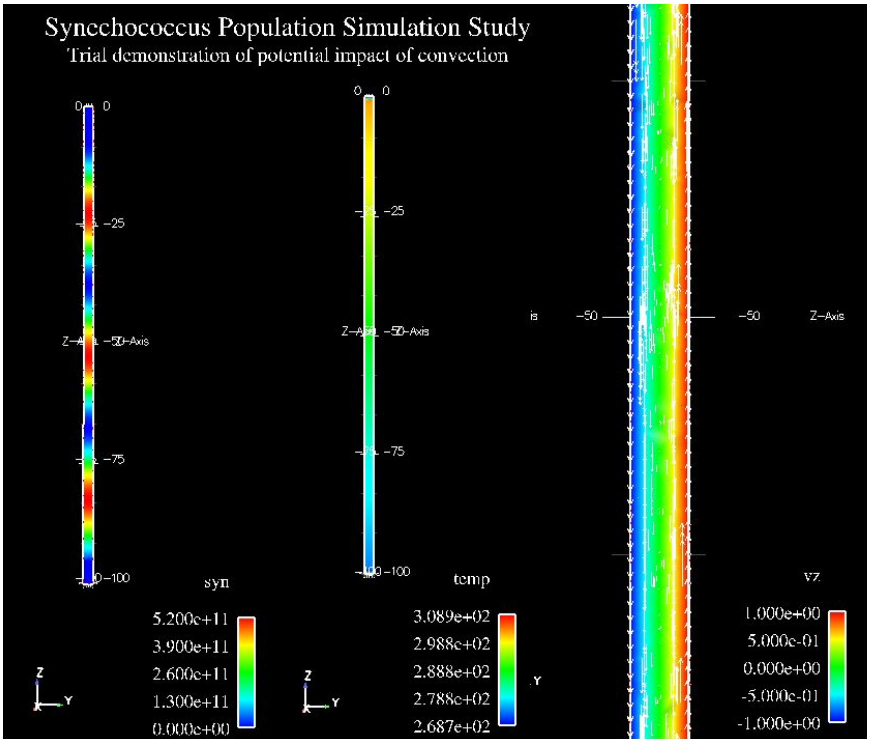


Figure 4-4. Effect of convection on microbe concentration and water temperature (respectively), using the given velocity profile.

<u>*Hierarchical Simulation Platform*</u>: We completed many of the tasks reported in the last Quarterly Report's 'Progress Towards Milestones' by synthesizing the work of a number of oceanographic experts modeling nutrient concentrations. This includes modeling the relationship between the rate of uptake of extracellular nutrients and intracellular quotas. Following a recent development by Grover (2003), we describe V_t^j as the cellular rate of uptake of the oceanic nutrients carbon, nitrogen, and phosphorous for $j \in \{C, N, P\}$ at time *t* with the following equation:

$$V_t^j = V_{max}^j \left(\frac{[j]_t}{K_j + [j]_t} \right) \left(\frac{Q_{max}^j - Q_t^j}{Q_{max}^j - Q_{min}^j} \right) e^{E_a^V / kT_t}$$

V^j_{max} is the maximal uptake rate of the nutrient *j*. The second term on the right-hand side is a Michaelis-Menten function of extracellular nutrient concentration, where j_t is the nutrient concentration outside the cell at time *t*, and K_j is the half-saturation constant. The third term describes the maximal and the minimal cell quotas Q^j_{max}, Q^j_{min} respectively, where $Q^j_t = Q^j_{t-1} + V^j_{t-1} \Delta t$ relates the current cell quota to the rate of uptake. E^v_a is the activation energy for nutrient uptake, *k* is the Boltzmann constant, and T_t is the temperature at time *t*.

We then map the intracellular nutrient quotas to the rate of carbon fixation occurring in the carboxysomes. We calculate the photosynthetic rate (P_t) at time *t* using the cell quotas (Q_t^C, Q_t^N, Q_t^P) inside the cell via:

$$P_t \propto P_{max} \frac{\frac{Q_t^N}{Q_t^C} - \left(\frac{Q^N}{Q^C}\right)_{min}}{\left(\frac{Q^N}{Q^C}\right)_{max} - \left(\frac{Q^N}{Q^C}\right)_{min}} \left(1 - \frac{Q_{min}^P}{Q_t^P}\right) \min\left\{ \frac{E}{K_{1/2}^E + E}, \frac{[C]}{K_{1/2}^R + [C]} e^{E_a^R / KT_t} \right\} M_t^{3/4}$$

The first term on the right-hand side (P_{max}) is the maximal photosynthetic rate under light and nutrient saturated conditions. This is attenuated by the cell quotas following Geider, et al. (1998) and Grover (2003). The minimizing function on the bracketed terms captures the light- and RuBisCO-limiting dependency of the light and dark cycles respectively. In the light-limited domain, the rate is dominated by the irradiance *(E)* and is relatively insensitive to temperature. In the light-saturated domain, the rate is predominately limited by the rate at which RuBisCO fixes C, and this is described by a Michaelis-Menten function with a temperature dependency captured by the Boltzmann factor for the reaction (Gillooly, et al. [2001]). The rate of photosynthesis scales with the number of carboxysomes and body mass, and we capture that by noting that $P \propto B \propto M_t^{3/4} \propto \frac{dM}{dt}$ (West, et al. [2001], Enquist and Niklas [2001]).

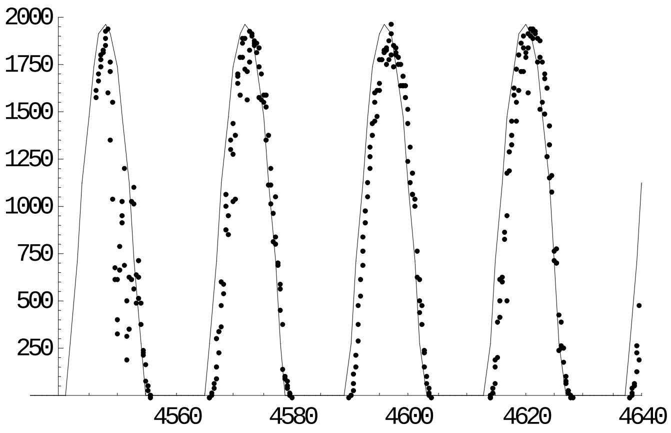
Progress Towards Milestones

Aim 1. Apply algorithms to Yeast proteome and publish and release scale-free network algorithms (9/04). The paper we submitted to *Bioinformatics* includes three inferences and comparisons of the Yeast protein network using phage display and two hybrid experiments. As detailed in our previous report, algorithms to infer scale-free networks are completed and have been tested.

Aim 2. Develop massively parallel versions of codes (6/04) and **Focus work on *Synechococcus* pathways associated with carbon sequestration (6/04).** Both of these milestones have been met and we have started working more closely with the experimentalists to focus on *Synechococcus* pathways.

Aim 3. This aim is currently on track with no immediate milestones. The focus during FY04 is on connecting the model as closely as possible with the biological problems of interest, and this is being done through the connections with the hierarchical modeling aspects of the problem.

Aim 4. Second iteration prototype of hierarchical modeling simulation (12/04) Concurrent with the above work at the cellular level, we are making progress towards this milestone by examining the appropriateness of Litchman and Klausmeier (2001) for modeling changes in irradiance over time at the ecosystem level. Below is a graph that shows the fit of Litchman and Klausmeier (2001) to data from the Hawaiian Ocean Time-series (HOT 2001) data:



Abscissa is hours since January 1, 2001 [domain: Cruise 128: July 9 – 13]; ordinate is irradiance in units of $\mu E m^{-2} s^{-1}$. The fit is representative of all 12 cruises sampling the year 2001. We find this a promising direction and expect to incorporate it in the model in the upcoming quarter. Additionally we expect to extend the application of this work by combining it with Beer's Law to account for the significant attenuation of light in the water column.

Collaboration With Others

The hierarchical modeling work has been incorporated into a newly submitted Sandia LDRD proposal led by Mark Boslough entitled "Coupling Biogeochemistry into Climate Models with Macroecological Scaling." This quarter also marked the first meetings focused on the details of coupling the hierarchal simulations with the continuum simulations.

Publications and Presentations

Faulon, J.-L., Martin, S. "Dynamical Robustness in Gene Regulatory Networks." C*onference Paper, IEEE CSB 2004*, Accepted. 2004.

Martin, S., Davidson, G., May, E., Werner-Washburne, M., Faulon, J.-L. "Inferring Genetic Networks from Microarray Data." *Conference Paper, IEEE CSB 2004*, Accepted. 2004.

Martin, S., Roe, D., Faulon, J.-L. "Predicting Protein-Protein Interactions Using Signature Products." *Bioinformatics*, Submitted. 2004.

Plimpton, S. J. "A Particle-Based Cellular Model of Protein Interactions." *SIAM Conference on Parallel Processing for Scientific Computing.* San Francisco, CA. February 2004.

Plimpton, S. J. "ChemCell: A Particle-Based Cellular Model of Protein Interactions." *Presentation at External Review of Sandia's Biotechnology Program*. Albuquerque, NM. February 2004.

Plimpton, S. J. "ChemCell: A Cellular Response Model for *Synechococcus.*" *Presentation at GTL External Advisory Board Meeting*. Phoenix, AZ. March 2004.

Slepoy, A. “ChemCell: Three-dimensional Stochastic Simulation of Bio-molecular Pathways.” *2nd New Mexico Workshop on Computational Cell Biology*. Santa Fe, NM. January 2004.

Reference

Enquist, B. J., Niklas, K. J. “Invariant Scaling Relations Across Tree-Dominated Communities.” *Nature,* 410: 655-660. 2001.

Geider, R. J., MacIntyre, H. L., Kana, T. M. “A Dynamic Regulatory Model of Phytoplankton Acclimation to Light, Nutrients, and Temperature.” *Limnol. Oceanogr*. 43: 679-694. 1998.

Gillooly, J. F., Brown, J. H., West, G. B., Savage, V. M., Charnov, E. L. “Effects of Size and Temperature on Metabolic Rate.” *Science* 293: 2248-2251. 2001.

Grover, J. P. “The Impact of Variable Stoichiometry on Predator-Prey Interactions: a Multinutrient Approach. *Am. Nat.* 162: 29-43. 2003.

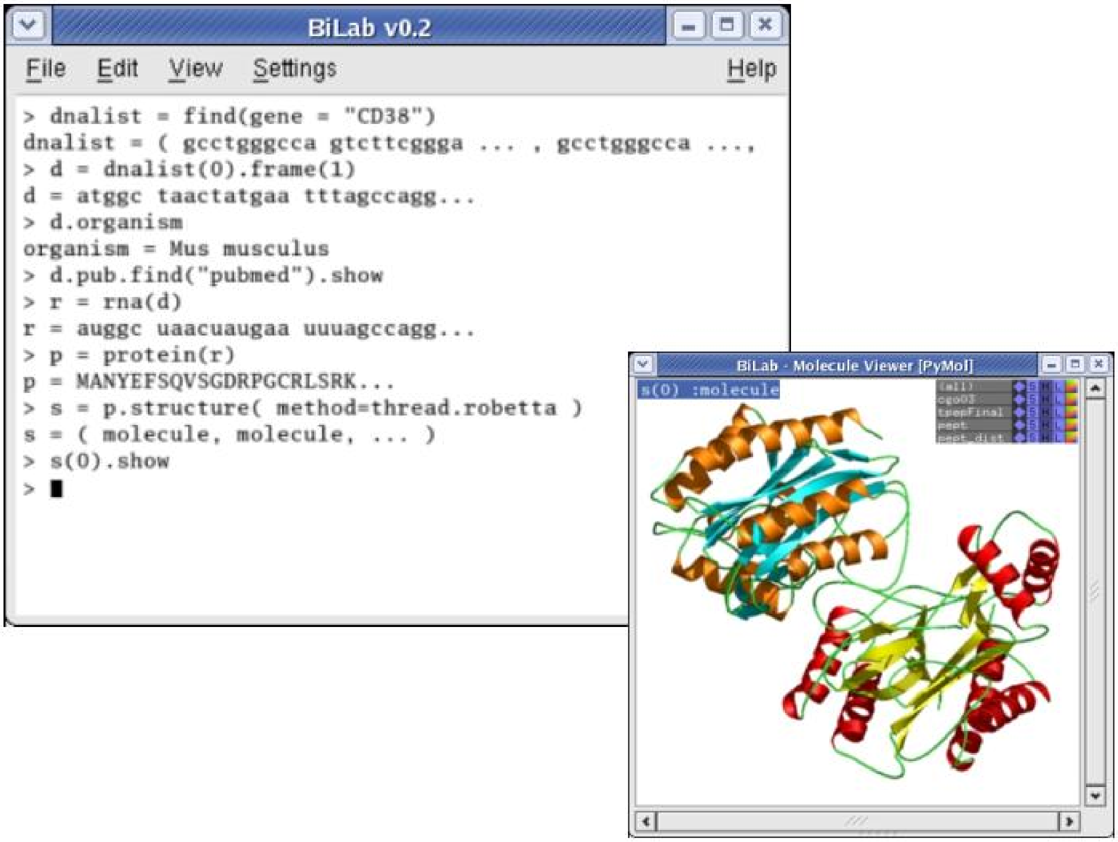
HOT (Hawaiian Ocean Time-series) 2001. Available at: http://hahana.soest.hawaii.edu/hot/hot-dogs/interface.html.

Jansen, R., et al. “A Bayesian Networks Approach to Predicting Protein-Protein Interactions from Genomic Data.” *Science*. 2003.

Litchman, E., Klausmeier, C. A. “Competition of Phytoplankton Under Fluctuating Light.” *Amer. Nat*., 157: 170-187. 2001.

West. G. B., Brown, J. H., Enquist, B. J. “A General Model for Onto Genetic Growth.” *Nature* 413:628-631. 2001.

Subproject 5: Computational Biology Work Environments and Infrastructure



```
> dnalist = find(gene = "CD38")
dnalist = ( gcctgggcca gtcttcggga ... , gcctgggcca ...,
> d = dnalist(0).frame(1)
d = atggc taactatgaa tttagccagg...
> d.organism
organism = Mus musculus
> d.pub.find("pubmed").show
> r = rna(d)
r = auggc uaacuaugaa uuuagccagg...
> p = protein(r)
p = MANYEFSQVSGDRPGCRLSRK...
> s = p.structure( method=thread.robetta )
s = ( molecule, molecule, ... )
> s(0).show
> █
```

Figure 5-1. We are creating a MATLAB^{R}-like work environment tailored to bioinformatics. The prototype showing the MATLAB-like command interface executing on biological objects rather than mathematical objects is shown above. Output from the tool can also be displayed visually as shown in the attached window.

Our computational work environments and infrastructure effort made significant progress this quarter toward establishing a data and computational infrastructure for use by the project. The *Synechococcus* Encyclopedia became available for all project members to use. New search capabilities were added to the Encyclopedia based on user feedback. Work continued on the MATLAB^{R}-like biology tool, “BiLab.” A significant accomplishment this quarter was the completion of the scigol language upon which BiLab is based and incorporation of all the biojava functions into BiLab. We developed and installed the Data Entry and Browsing tool to capture the experiments’ metadata in various phases of the project, from microbe cultures to production of microarrays and their analysis, as well as the experiments to identify the cellular functions.

This work was funded in part or in full by the US 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).

Accomplishments

Our computational work environments and infrastructure effort made significant progress this quarter toward establishing a data and computational infrastructure for use by the project.

Synechococcus Encyclopedia: Several groups have been generating data based on their experimental and computational investigations of *Synechococcus* Sp. This data is heterogeneous and widely spread across various institutions. To facilitate integrative, exploratory, and in-depth analysis of this organism, we designed a centralized knowledge-base system that will enable a semantically and syntactically consistent view of all these data and complex queries across multiple data types. The data sources that are a part of the knowledge-base system include protein- and RNA-coding genes, Pfam domains, Blocks motifs, InterPro signatures, COG, and KEGG-based functional assignments. At the structure level, they consist of secondary and tertiary structural models predicted by PROSPECT and other tools, SCOP-based functional assignments, and FSSP profiles for homologous protein sequences. Microarray data has begun to be stored on the data server as well. Accomplishments this quarter on the data server include the following.

- Implemented the data integration and exploration framework. This framework supports the integration of the heterogeneous data. New data sources can be added to the framework with very little effort.
- Designed the schema for the available data sources and uploaded them into the relational database.
- Implemented the Query generator/interpreter to represent user queries and retrieve appropriate results from the relational database.
- Designed and implemented the front-end interface to the system, by which the integrated data can be explored using simple and complex queries.
- Designed and implemented Keyword Search capability to browse the metadata of the knowledge-base to aid query formulation.

The *Synechococcus* Encyclopedia will ultimately comprise information at many levels: genome sequence, structure, regulation data, protein interactions, systems biology, and experimental microarray data.

BiLab MATLAB-like Biology Tool: We made significant progress in developing BiLab, a MATLAB™-like work environment tailored to bioinformatics. Because MATLAB is a proprietary language not available for modification, a new language, named *scigol*, which is targeted at scientific computing and bioinformatics in particular, has been designed. The syntax is similar to MATLAB, but more powerful. Scigol forms the basis for interactive evaluation of expressions in a command-line environment called BiLab. Work is currently ongoing to add more biological data-types, bioinformatics APIs, and to integrate visualization and analysis tools into BiLab.

The scigol language has been designed to bridge the gap between traditional languages used for scientific computing, such as Fortran, C, and C++, which are statically typed and consequently high-performance, and the trend in bioinformatics to use newer scripting languages, like Perl and Python, which are dynamically typed. Scigol has been designed with both static and dynamic typing; weak typing for rapid prototyping, exploratory programming and scripting, and strong typing for support of larger software codes. Scigol incorporates all the features expected in modern programming languages, like C# and C++. Scigol was also designed with ease-of-use and a shallow learning curve in mind. It is suitable for interactive use (in BiLab), where the more advanced features of the language – classes, namespaces, first class types and functions, rich operators, design-by-contract, and others – can be selectively ignored.

One characteristic of the current bioinformatics landscape is a proliferation of distinct tools and codes written in a variety of languages for various platforms. Hence, one requirement for BiLab is to be able to interact with or incorporate existing codes in other languages. This has enabled the integration of the bioJava API into BiLab, upon which some of the basic biological data-types are being built. Work is underway to incorporate the bioperl and biopython API libraries as well. It is also planned to enable easy access to external libraries in Fortran, C, and, time permitting, C++.

The next step in the implementation of BiLab will be to start incorporating more visualization and analysis tools. There are a large number available, from which a couple will initially be chosen for integration. Example visualization packages include NCBI tools**,** OpenRasmol, ClustalX, Cn3D, Garlic, and OpenBabel.

The completion of the BiLab language will pave the way for the implementation and integration of various bioinformatics tools as hooks used by the runtime to manipulate biological data-types, such as chemical reactions, proteins, and gene sequences.

DEB: Data Entry and Browsing Tool: The DEB system was developed with the guidance of the biologists on the project as a tool to capture the experiments' metadata in various phases of the project, from microbe cultures to production of microarrays and their analysis, as well as the experiments to identify the cellular functions. The technology we use is the development of user interfaces for data entry and browsing based on schema descriptions. To achieve this efficiently, we take advantage of an object-based system that provides an object implementation on top of the Oracle database system, called OPM. The DEB system is built on top of the OPM system. The main accomplishments during the last quarter are described below.

Developed a schema-driven user interface tool.
The DEB tool was designed to be generic, in that the same type of interface can be used for different purposes. To achieve this, we developed a User Interface Generator (called DEB-UIG) that takes a description of the schema and generates web-based interfaces automatically from that schema. In order to accomplish this goal, every element of the schema is "decorated" with a User-Interface type that determines how that element will be presented on the screen and what functions can be performed with it. The DEB-UIG was fully implemented.

Developed a schema and a DEB interface for Nucleotide Pools.
The definition of the schema was started in November of 2003, and further refined in December 2003 in several face-to-face meetings with Brian Palenik and his assistant, Lori Crumbliss, at SDSC. Over the last quarter, a DEB interface was generated and finalized. Lori has entered 24 instances of the relevant metadata of the microbe cultures and conditions in which they were generated into the database. This is now available on-line.

Added features to upload and link files.
In the previous version of DEB it was only possible to upload a file from the user's system, such as .doc .pdf and .xls files. However, in GTL, we also have large files that are loaded separately into the system at ORNL. We decided to add a feature that allows the DEB metadatabase to be linked to these files without the need to upload them. This required adding a new generic object type, called file references into all the schemas by default. This feature was fully developed.

Added features to keep history of file sources and modifications.
Tony Martino, who was the main biologist guiding the design of the DEB interfaces, pointed out that for use in his lab, it will be necessary to have the source of the information loaded in the DEB database. If the main ORNL site is not available (temporary maintenance, etc.) it will be good to have a reference to the original copy as well as the complete data in DEB. To accommodate that, we have developed a report generation feature, called DEB-REP that generates an HTML page with the entire history of the data entry as well as the source of uploaded files. In addition, we added "person last modified" and "last-entered" elements to each entry, including uploaded files and DEB entries, so that in case of modifications, the information on the change is captured.

Our next goal is to work with Tony Martino on using the DEB tool in his lab. A schema was already designed and generated for his lab. In addition, we started to work with Ian Paulsen from TIGR on defining the schema that will capture the metadata from his lab into DEB. That schema will be designed to be compliant with the MIAME standard guidelines.

Web site and Electronic Notebook: Our Subproject 5 team continues to maintain the project's web site, keeping calendars, publications, and highlights updated. We also updated the project's Electronic Notebook software to provide new upload features.

Progress Towards Milestones

Aim 1. Integrate new methods and tools into an easy-to-use work environment. (2/04). In meeting this Aim, we have identified *Synechococcus* dataserver requirements and identified and collaborated with various groups producing the data to be deposited into the data server. We have installed the necessary compilers and support libraries to accommodate the OPM and DEB data collection software on the *Synechococcus* data server.

Aim 2. Complete design of general-purpose, graph-based data management system (2/04). There is no new information to report this quarter

Aim 3. Develop efficient data organization and processing of microarray databases by setting up a database using a standard MIAME scheme (2/04). DEB is implemented and data has begun to be input by experimentalists. Also, DEB is being ported to a SUN server hooked to the *Synechococcus* Encyclopedia.

Aim 4. Develop new cluster analysis algorithms, specifically a new tool that can do cluster analysis across multiple genomic databases around the country (9/03). This Aim has been accomplished.

Aim 5. Set up environment that allows researchers in this proposal to utilize the computational resources at ORNL and Sandia (2/04). To satisfy this Aim, we have set up the ModPod data server with firewall exceptions to allow web access. Automatic backups are in place with daily incremental and weekly full backups. We have applied all security patches and updates to the ModPod data server. The Scigol language has been developed for the MATLAB-like bioinformatics work environment. A new java-based electronic notebook version was developed for the project.

Collaboration With Others

The *Synechococcus* Encyclopedia is a very collaborative effort that is highly supported by a number of groups within the Center as well as other GTL Centers, including Frank Larimer and

Mariam Land, Ying Xu and Serguei Passovets, and Jerilyn Timlin. We started a collaboration with Adam Arkin on utilizing our Encyclopedia infrastructure for hosting data on multiple genomes available at Arkin's lab.

In developing the DEB tool, we held several face-to-face meetings with Brian Palenik and his assistant, Lori Crumbliss, at SDSC. Tony Martino was the main biologist guiding the design of the DEB interfaces. Additionally, we started to work with Ian Paulsen from TIGR on defining the schema that will capture the metadata from his lab.

Publications and Presentations

Samatova, N. F., Geist, A., Chandramohan, P., Krishnamurthy, R., Yu, G-X., Heffelfinger, G. "The *Synechococcus* Encyclopedia." *2004 DOE Genomes:GTL Grantees Workshop*, Abstract and Poster Presentation. Washington, D.C. March 1-2, 2004.

Samatova, N. F., Geist, A., Chandramohan, P., Krishnamurthy, R., Yu, G-X., Heffelfinger, G. "The *Synechococcus* Encyclopedia." *2004 DOE Genomes:GTL Grantees Workshop*, Demonstration. Washington, D.C. March 1-2, 2004.

DISTRIBUTION FOR JUNE 2004 GTL QUARTERLY SAND REPORT

SAND2004-5762

10	MS-0885	Grant Heffelfinger, 1802
1	MS-9018	Central Technical Files, 8945-1
2	MS-0899	Technical Libr6ary, 9616