

Genomes to Life Project Quarterly Report April 2005

Grant S. Heffelfinger
Biological and Energy Sciences Center
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-1413

Abstract

This SAND report provides the technical progress through April 2005 of the Sandia-led project, "Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling," funded by the DOE Office of Science Genomics:GTL Program.

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

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

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

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

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

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

More information about this project can be found at www.genomes-to-life.org

Acknowledgment

We want to gratefully acknowledge the contributions of:

Grant Heffelfinger^{1*}, Anthony Martino², Brian Palenik⁶, Andrey Gorin³, Ying Xu^{10,3}, Mark Daniel Rintoul¹, Al Geist³, Matthew Ennis¹, with Pratul Agrawal³, Hashim Al-Hashimi⁸, Andrea Belgrano¹², Mike Brown¹, Xin Chen⁹, Paul Crozier¹, PguongAn Dam¹⁰, Jean-Loup Faulon², Damian Gessler¹², David Haaland¹, Victor Havin⁴, C.F. Huang⁵, Tao Jiang⁹, Howland Jones¹, David Jung³, Katherine Kang¹⁴, Michael Langston¹⁵, Shawn Martin¹, Shawn Means¹, Vijaya Natarajan⁴, Roy Nielson⁵, Frank Olken⁴, Victor Olman¹⁰, Ian Paulsen¹⁴, Steve Plimpton¹, Andreas Reichsteiner⁵, Nagiza Samatova³, Arie Shoshani⁴, Michael Sinclair¹, Alex Slepoy¹, Shawn Stevens⁸, Charlie Strauss⁵, Zhengchang Su¹⁰, Ed Thomas¹, Jerilyn Timlin¹, WimVermaas¹³, Xiufeng Wan¹¹, HongWei Wu¹⁰, Dong Xu¹¹, Grover Yip⁸, Erik Zuiderweg⁸

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

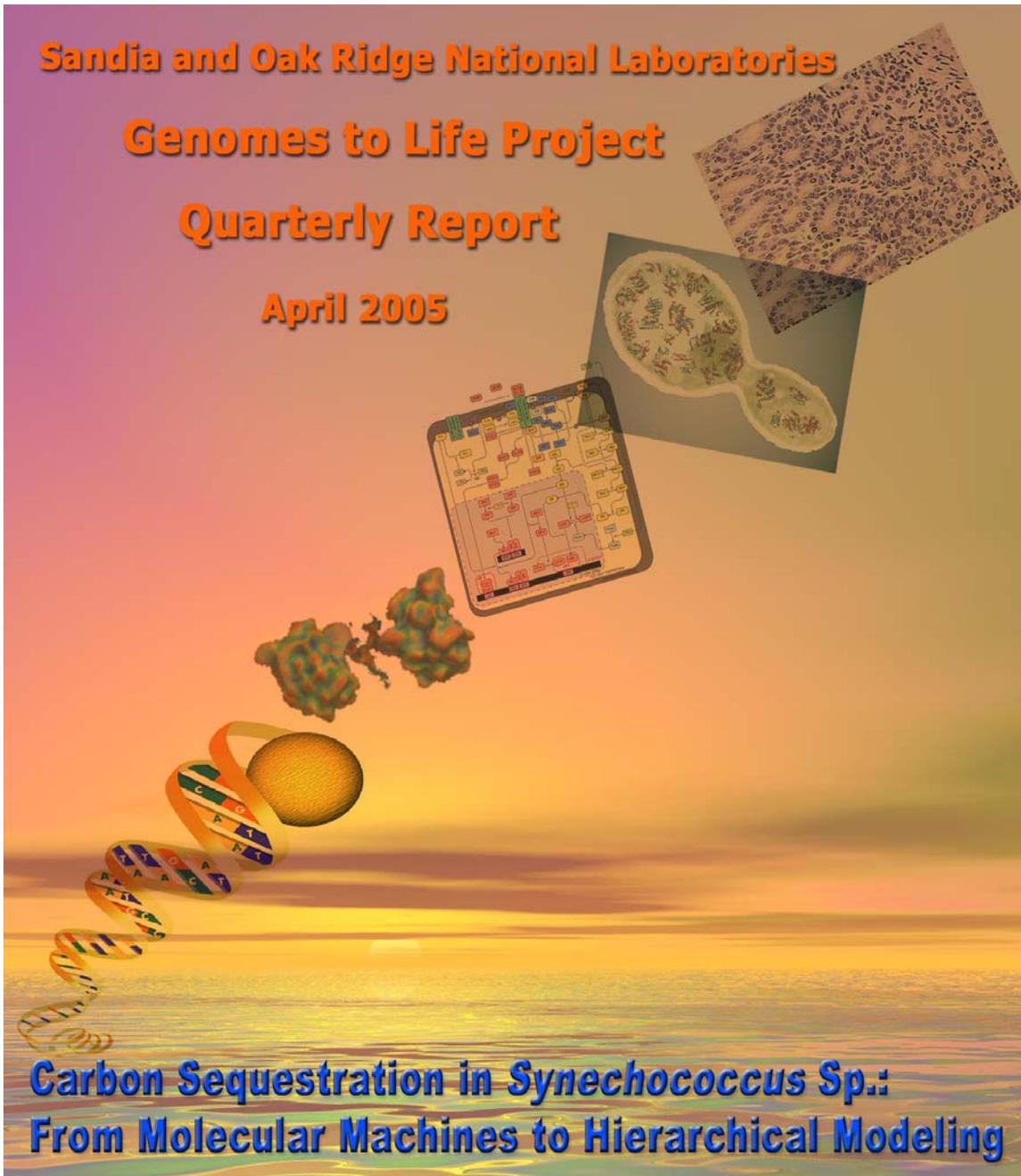
1. Sandia National Laboratories, Albuquerque, NM
2. Sandia National Laboratories, Livermore, CA
3. Oak Ridge National Laboratory, Oak Ridge, TN
4. Lawrence Berkeley National Laboratory, Berkeley, CA
5. Los Alamos National Laboratory, Los Alamos, NM
6. University of California, San Diego
7. University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

Sandia and Oak Ridge National Laboratories

Genomes to Life Project

Quarterly Report

April 2005



**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-AC04-94AL8500.

Table of Contents

| | |
|--|----|
| Table of Contents | 6 |
| Executive Summary | 7 |
| De Novo Method in Structure Prediction and Mass Spec | 8 |
| Hyperspectral Imaging of <i>Synechocystis</i> | 9 |
| Robust Statistical Analysis of Microarray Data with Replicated Spots | 11 |
| Proteomic Toolshop | 13 |
| Protein Interaction Network Inference | 14 |
| Computational Inference of Biological Networks in Cyanobacteria Genomes..... | 15 |
| NMR Studies in RuBisCO..... | 17 |
| Biopathways Graph Data Management System | 20 |
| Data Entry and Browsing (DEB) Tool | 20 |
| Event-driven, Hierarchical Modeling of <i>Synechococcus</i> Dynamics | 22 |
| Molecular Modeling of RuBisCO's Gating Mechanism | 24 |
| Spatial Modeling of <i>Synechococcus</i> Ecosystems | 25 |
| Spatio-temporal Cell Modeling with Particles | 26 |
| Publications | 28 |
| Presentations..... | 30 |

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

Executive Summary

We report significant progress this quarter in the 3-D hyperspectral confocal fluorescence imaging of wild-type and mutants of the *Synechocystis* bacteria. Our 3-D hyperspectral microscope came on line in November 2004, and is now operating at diffraction-limited spatial resolution in the 490-800 nm spectral range. These studies are in collaboration with Prof. Vermaas at Arizona State University (ASU), who is independently funded by DOE's Microbial Cell Project to investigate the photosynthetic process in cyanobacteria.

Computationally, we report the development of a fully automated, structure-based annotation pipeline which we applied to the annotation of hypothetical sequences in *Synechococcus*. We also carried out one of the first very large-scale comparisons spanning such a large function space and using a broad-based gold standard for comparison demonstrating that roughly forty percent of sequences in pFam can be assigned uniquely to their correct family based on literature retrieval. We also discuss the development of a suite of computational prediction tools in support of comparative genome analysis and inference of biological networks, including P-MAP, JPOP, Excavator, CUBIC, PRIME, PROSPECT-PSPP and Promoco. Our efforts this quarter toward developing computational tools to infer protein interaction networks involved applying our signature support vector method for the prediction of β -sheet topology demonstrating 80-90% accuracy for the entire PDB database. After our signature support vector paper appeared in *Bioinformatics*, we immediately received more than ten requests for our support vector code as well as requests for the code from various biotech and pharmaceutical companies, including AstraZeneca, Genentech, MDL, Daylight, and IntelliChem.

We completed the development of an accelerated version of our LAMMPS molecular dynamics software that has enabled the simulation of the gating mechanism for tobacco, rice, and *Synechococcus* WT RuBisCOs as well as several mutant forms of *Synechococcus* RuBisCO. Our progress this quarter on our cell simulation tool involved implementing additional reaction algorithms within the ChemCell framework including both the nonspatial Gillespie SSA as well as a new spatial form of the SSA. We also performed new simulations of the carbon-fixation pathway model for *Synechococcus* demonstrating that 1) glucose production was reduced by a factor of 2 when carbonic anhydrase (CA) is only present in the cytoplasm versus the carboxysome, and 2) glucose production was increased by 30-50% when the CA was layered at the carboxysome surface. These qualitative results indicate how ChemCell can be used to include spatial realism in a pathway model. We also report on our efforts to combine nutrient stoichiometry and metabolic rate allometry in a single, unified dynamic model of marine carbon sequestration including a new simulation approach for event-driven, disparate-time hierarchical modeling. We report computational evidence which strongly support the Energy Equivalency Rule, which shows that as populations consume more carbon faster with increased temperature, the total carbon fixation rate for the community remains constant.

Finally, work also continues on our computational infrastructure tools including our proteomic toolshop codenamed BiLab, a MATLAB[®]-like tool that understands and can operate on biological objects. Our most recent accomplishments include the incorporation of the ability to read in data directly from external biological databases and put them into BiLab datatypes that allow analysis and manipulation by all BiLab functions. Our web-based Data Entry and Browsing (DEB) tool, designed to facilitate capturing the metadata from experiments and laboratories and store them in a database in a computer searchable form also reports substantial progress this quarter with new schemas were developed for the TIGR hybridization metadata and Sandia's hyperspectral imaging metadata enabling a large increase in the number of data entries.

De Novo Method in Structure Prediction and Mass Spec

Los Alamos National Laboratory

Charlie Strauss, Andreas Reichsteiner, C. F. Huang, Roy Nielson, Andrey Gorin

As part of Subproject 2, Computational Discovery and Functional Characterization of *Synechococcus* Sp. Molecular Machines in the Sandia National Laboratories-Oak Ridge National Laboratory Genomics:GTL project “Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling” we are working on several ongoing activities, including:

- Automated literature mining for confirmation of structural annotation;
- Development of a portable version of *Robetta*;
- Markov assembly of *de Novo* mass spectra;
- Processing of *Synechococcus* targets through *Robetta* and *Mammoth*;
- General advances in *Rosetta* and *Mammoth* algorithms; and
- Scalable search algorithm development for improved sampling of conformational space

Development of *de Novo* mass spectra is a core part of the *Synechococcus* Genomics:GTL effort. The *Robetta* server is both a major product, as a public server, and a tool. Adding automated literature analysis to the structure-based annotation will help us overcome what is, ironically, the major deficiency in the usefulness of the product: it produces more annotations than can be checked by hand. Because the annotations it produces are, intentionally, remote homology based on structure, confirmation of any prediction is significantly more laborious than traditional sequence annotation, hence the desire to automate some of this process. The popularity of the public server caught us by surprise and we sought to produce a portable version. Work on that was shut down at LANL for the last six months following the LANL-wide shut down. We recently have acquired a technician to resume work on this).

Looking at the big picture, over the course of this project we have developed a fully automated, structure-based annotation pipeline and have applied it to the annotation of hypothetical sequences in the target organism.

Some significant recent results include the *de Novo* mass spectra Markov model rewritten to allow higher order, conditional interactions. This should allow increased use of prior knowledge probabilities. We have also ported the original algorithm from Perl to Python for better development and easier communication of results. Demonstration Medical Subject Heading terms in literature can be compared to find similarities between proteins and are sufficient to classify protein function at the level of pFAM, without any use of sequence comparison. This is one of the first very large-scale comparisons spanning such a large function space and using a broad-based gold standard for comparison. Roughly forty percent of sequences in pFam can be assigned uniquely to their correct family based on literature retrieval. About 75 percent can be assigned ambiguously to one of five possible families, usually related in function. We have also published a simplified chain folding algorithm and presented research results at conferences.

de Novo mass spectra is of crucial interest to the mass spec farms at Oak Ridge and PNNL. Structure analysis is of broad interest to the characterization of unique genes in genomes. Many Genomics:GTL projects either focus on the unique specialized capabilities of a microbe or are attempting broad discovery of new genomes (e.g., Venter Institute). In both cases the unique features of the genome are hardest to annotate by traditional sequence based methods. Other projects peripherally interacting with Genomics:GTL projects, such as the integrated pathway analysis at the Institute for Systems Biology, have used *Robetta* to fill in missing pathway elements.

Robetta has a worldwide impact with sequences being submitted from major institutions around the world. The backlog can run as high as 30 to 60 days. *de Novo* mass spectra will someday unleash the equivalent to proteomics as sequencing is to genomics.

Hyperspectral Imaging of *Synechocystis*

Sandia National Laboratories

David Haaland, Jerilyn Timlin, Howland Jones, Michael Sinclair, and Wim Vermaas
(*Arizona State University*)

During this quarter, we have made significant progress in the 3-D hyperspectral confocal fluorescence imaging of wild-type and mutants of the *Synechocystis* bacteria. Our 3-D hyperspectral microscope came on line in November 2004, and is now operating at diffraction-limited spatial resolution in the 490-800 nm spectral range. These studies are in collaboration with Prof. Vermaas at Arizona State University (ASU), who is independently funded by DOE's Microbial Cell Project to investigate the photosynthetic process in cyanobacteria. Cyanobacteria have a number of photosynthetic pigments that cannot be imaged and quantitatively separated using commercial fluorescence microscopes due to their high degree of spectral and spatial overlap. Our new hyperspectral imaging microscope, coupled with Sandia's proprietary multivariate curve resolution (MCR) software, can extract spatial and quantitative relative concentration information for all the pigments in the cell. Thus, we can recover the pure emission spectra of phycocyanin, protochlorophyllide, allophycocyanin, allophycocyanin B, and chlorophyll, which all have emission maxima in the spectral region from 630 to 684 nm. Simultaneously, we are able to obtain quantitative concentration maps of each pigment in three dimensions with an x-y spatial resolution of 250 nm and a z spatial resolution of ~600 nm. Our hyperspectral data provide the first visualization of the heterogeneous distributions of these pigments both between and within the cells.

This work is part of both Subproject 1 (Experimental Elucidation of Molecular Machines and Regulatory Networks in *Synechococcus* Sp.) and Subproject 3 (Computational Methods Towards The Genome-Scale Characterization of *Synechococcus* Sp. Regulatory Pathways) in the Sandia National Laboratories-Oak Ridge National Laboratory Genomics:GTL project "Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling," project and is relevant to *Synechococcus* Genomics:GTL (SGTL) since the photosynthetic process in cyanobacteria eventually leads to carbon sequestration and *Synechocystis*, the first photosynthetic organism whose genome was completely sequenced, is a freshwater cousin of *Synechococcus*. By exploiting the library of targeted gene deletion mutants that are available, we can study the photosynthesis and electron transport pathways in cyanobacteria. This project also positions us to achieve our hyperspectral imaging goals outlined in the Sandia-led FY06-FY10 Genomics:GTL proposal, "Bridging the Gap: From GTL Pipelines to Predictive Models."

Over the course of the project, we have applied 2-D hyperspectral imaging to the evaluation and understanding of microarray data and experimental problems that limit their reproducibility. Although we continue that effort, we have added the capability of confocal hyperspectral imaging in order to obtain quantitative 3-D images of multiple fluorescent components in live bacteria. This enabling technology, along with developments in MCR algorithms and user-friendly data visualization and analysis software, allows us to study processes in living cells that were not possible before the invention and practical implementation of these technologies.

In our hyperspectral imaging experiments, we find that the *Synechocystis chlL*- mutant, which is unable to convert protochlorophyllide to chlorophyll when grown under dark conditions, indeed lacks chlorophyll (the 684 nm emission component). Also, allophycocyanin B and protochlorophyllide emission are major spectral contributors in the cells. We observe that these pigments remain at the periphery of the cell, even though true thylakoid membranes are known to be absent in these mutants. However, when the same cells are grown in light, the protochlorophyllide is converted to chlorophyll and chlorophyll emission is observed. Some protochlorophyllide remains and becomes concentrated in small foci around the periphery of the cell. In a *Synechocystis PSI-less* mutant that lacks photosystem I, we observe that these cells only have 25 percent of the chlorophyll found in the wild type. The allophycocyanin B emission is still observed since not all allophycocyanin B is connected to chlorophyll. Generally, we find that chlorophyll fluorescence is most intense along the periphery, as expected since chlorophyll is associated with proteins bound to the thylakoid membranes, but chlorophyll fluorescence is also observed within the cell at higher relative concentration levels than the other pigments. Both phycobilin and protochlorophyllide fluorescence are primarily found along the cells' periphery, suggesting a distinct difference in localization of the various pigments in the cells.

For the remainder of the year, we will perform two more sets of imaging experiments with *Synechocystis* and various photosynthetic mutants of *Synechocystis* in order to confirm our original experiments and to provide further evidence for our assignments of the emission pigments. We will also study the photobleaching properties for these bacteria, since preliminary studies suggest a nonlinear photobleaching process that is selective for only one of the pigments. These results will also be reported in at least two journal articles.

The impact to the Genomics:GTL world outside of our project is that we will be able to team with other Genomics:GTL scientists to study protein-protein interactions in *Rhodospseudomonas palustris* (R. pal). These studies are currently not possible with commercial confocal fluorescence microscopes due to the strong autofluorescence of the photosynthetic pigments in R. pal. This work is already impacting the DOE-funded Microbial Cell Project through our collaboration with Professor Vermaas to study *Synechocystis*. Our ability to track and follow multiple proteins within a bacterial cell is critical for providing input for and validating predictions of computational models of photosynthetic bacteria. We are not limited to photosynthetic bacteria and the benefits of hyperspectral imaging can easily be realized in non-photosynthetic bacteria through fluorescent labeling of proteins. The impact beyond the Genomics:GTL program is that we have developed a technology that is becoming recognized as an enabling technology for biologists to study the spatial and temporal generation and behavior of multiple molecules in living cells that would not be possible with other microscopy techniques.

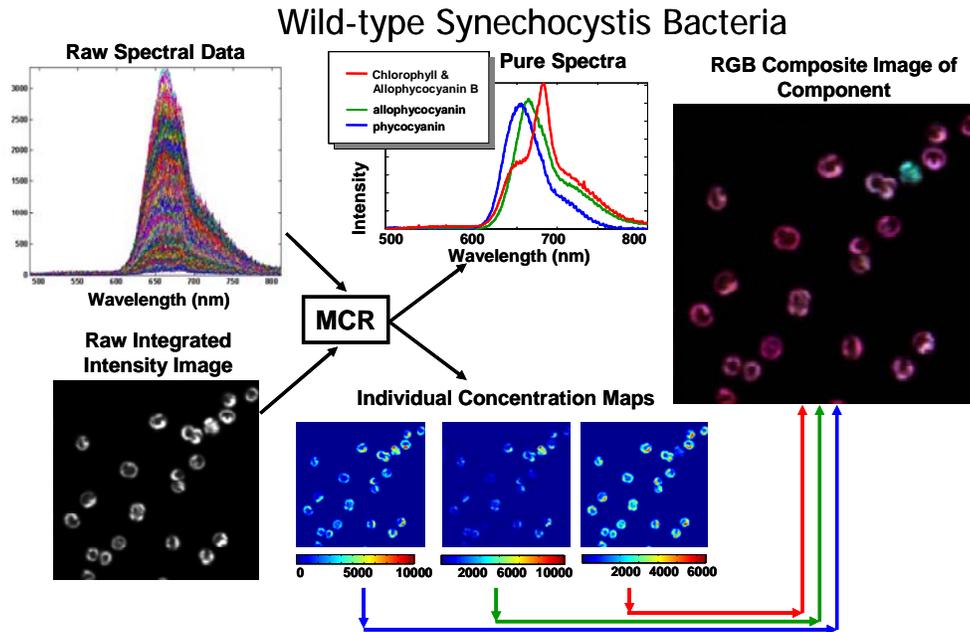


Figure 1: Hyperspectral *in vivo* image of Wild-type *Synechocystis* bacteria showing the spectra from a single z-slice of a 3D image. Pure emission spectra and relative concentration maps of the cells are shown along with a composite RGM image of the cells. The cyan cell in the 3D imaging is a dead cell.

Robust Statistical Analysis of Microarray Data with Replicated Spots

Ed Thomas¹, David Haaland¹, Katherine Kang², Brian Palenik³, Ian Paulsen², and Jerilyn Timlin¹

¹Sandia National Laboratories, ²The Institute for Genomic Research, ³Scripps Institution of Oceanography

As part of Subproject 3, Computational Methods Towards The Genome-Scale Characterization of *Synechococcus* Sp. Regulatory Pathways, in the Sandia National Laboratories-Oak Ridge National Laboratory Genomics:GTL project “Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling,” we have analyzed experimental data produced via a range of experiments that were performed in order to characterize the expression profiles of *Synechococcus* WH8102 in a wide variety of nutrient and stress conditions. The data were acquired from replicated dye-swap experiments with RNA extractions from multiple biological samples using a full *Synechococcus* genome microarray with replicated spots as the experimental platform. In addition to characterizing the biological response to the nutrient and stress conditions, these experiments were used to investigate and quantify sources of variation in these studies including spatial variation within an array as well as variation across biological samples and dye-related effects.

The Institute for Genomic Research’s (TIGR) SPOTFINDER and MIDAS software were used to process each microarray image. This processing resulted nominally in a data array consisting of the relative fluorescence intensities of each spot ($\frac{\text{Experimental}}{\text{Control}}$). Spots with low intensity were not automatically rejected, resulting in quantitative representation of a vast majority of the 2,531 genes over six spatially varying spots on each array. For each array, the expression of each

gene was summarized by the median relative intensity ($Y = \text{median}(\log_2(E/C))$) across its associated spots. This measure of gene expression is robust as it is insensitive to spot anomalies and/or large spatial effects, which were occasionally present. Array replication provided a direct mechanism to assess the reproducibility of gene expression across arrays.

The reproducibility of the gene expression across arrays within each experiment was assessed as follows. First, the gene expression data were normalized, by array, to remove global effects (across all spots on an array) associated with the dye and/or the biological replicate (see Figure 1). In general, following normalization, the reproducibility of the summary measure of gene expression across replicate arrays from the same biological sample was found to be excellent. Typically, for a given gene, the standard deviation of Y for replicates of the same biological sample (including dye swaps) was found to be about 0.15. However, the reproducibility of Y across biological samples was found to be more problematic. Currently, work is underway to understand the nature of this limiting source of variability.

For each experiment, the reproducibility of the measured gene expression across replicates was used to assess the statistical significance of the gene expression that was observed. The quality of a given experiment could easily be judged by the overall level of reproducibility within the experiment. Unusual experimental effects and anomalies that limit the ability to identify expressed genes were communicated both to the originators and users of the microarray data.

This work is relevant to *Synechococcus* Genomics:GTL (SGTL) because it supports the development of a fundamental understanding of *Synechococcus* regulatory networks via improved quality and reliability of experimental data and data summaries leading to more valid inference regarding gene expression. Also, it leads to error models that can be used in simulations to understand the effects of experimental variation on the process of constructing regulatory networks.

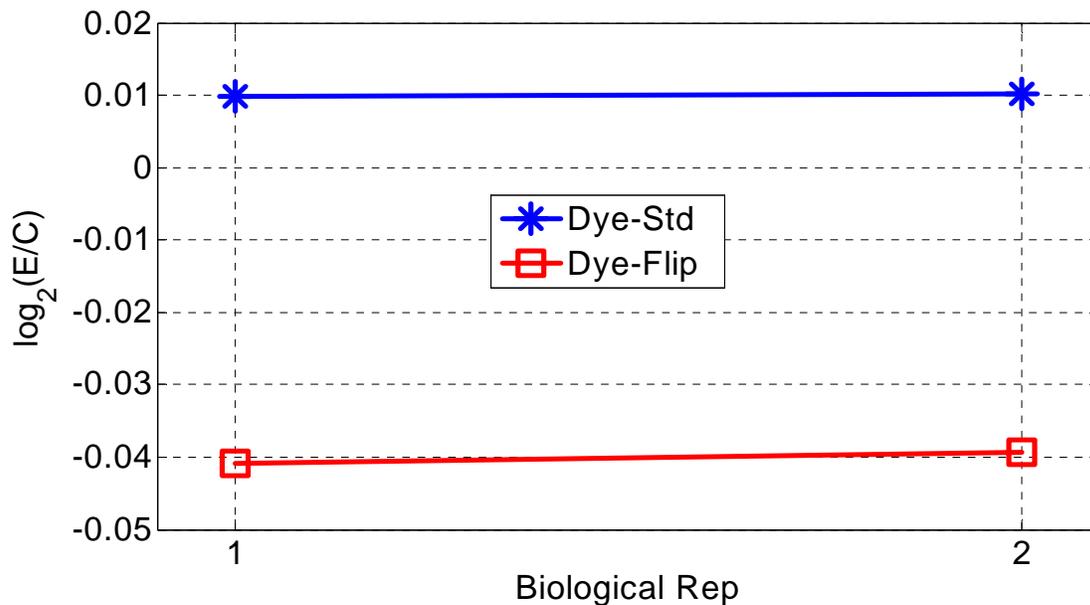


Figure 2: Global effects removed by normalization

Proteomic Toolshop

Oak Ridge National Laboratory

Al Geist, David Jung, and Pratul Agrawal

In FY 2004 we began development of a **Proteomic Toolshop** codenamed BiLab. The Toolshop is a MATLAB[®]-like tool that understands and can operate on biological objects. For example, one BiLab datatype is DNA and BiLab understands the operations that can be performed on DNA, such as transcribing it into another datatype called RNA (Figure 3, on following page). The initial prototype of the Toolshop can perform all the functions in the BioJava library and the National Center for Biotechnology Information (NCBI) tools library and understands a dozen biological datatypes. The Toolshop can also perform floating point matrix operations like MATLAB.

This research is part of Subproject 5, Computational Work Environments and Infrastructure, in the Sandia National Laboratories-Oak Ridge National Laboratory Genomics:GTL project “Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling.” Our contribution is significant to this Genomics:GTL project and computational biology in general because the biology community needs an easy-to-use tool that can unify the biological concepts and provide a simple environment where biologists can try out new ideas.

MATLAB transformed the numerical linear algebra community. Our goal is to do the same for the biological community by providing an interactive, intuitive language designed for bioinformatics. BiLab has a shallow learning curve, is fully featured to support large-scale toolbox development (i.e., support for strong typing, name-space control, generic types, Object-Oriented, Design-By-Contract, etc.), supports dynamic typing for interactive use and static typing for performance, and allows easy integration of existing bioinformatics tools and graphical toolkits and seamless access to all the data in online bioinformatics databases. The Toolshop provides consistent integration between command-line interaction and graphical interaction, which provides the best of both worlds and easy switching back and forth for the user.

Our most recent accomplishments include the incorporation of the ability to read in data directly from external biological databases and put them into BiLab datatypes that allow analysis and manipulation by all BiLab functions. This was demonstrated at the 2005 Genomics:GTL Principal Investigators meeting by manipulating data read in from GenBank. At that meeting, several biologists expressed interest in the Toolshop and Brian Palenik requested the ability to generate and operate on phylogenetic trees. Since the meeting, we have added “tree” as a new BiLab datatype and incorporated half a dozen functions from the Phylip phylogenetic tree package into BiLab.

In FY2005 we will make the Toolshop software more robust and portable so we can give it to researchers in the *Synechococcus* Genomics:GTL project for their use and for their feedback on the additional functions and biological datatypes required. The Systems Biology Workbench (SBW) team is interested in using BiLab as a front-end to their function library. This summer we will incorporate the SBW functions into the Toolshop. As time permits, we will investigate incorporating the protein function prediction, regulatory pathway prediction, and cell modeling applications from our project into Scigol and the Proteomic Toolshop.

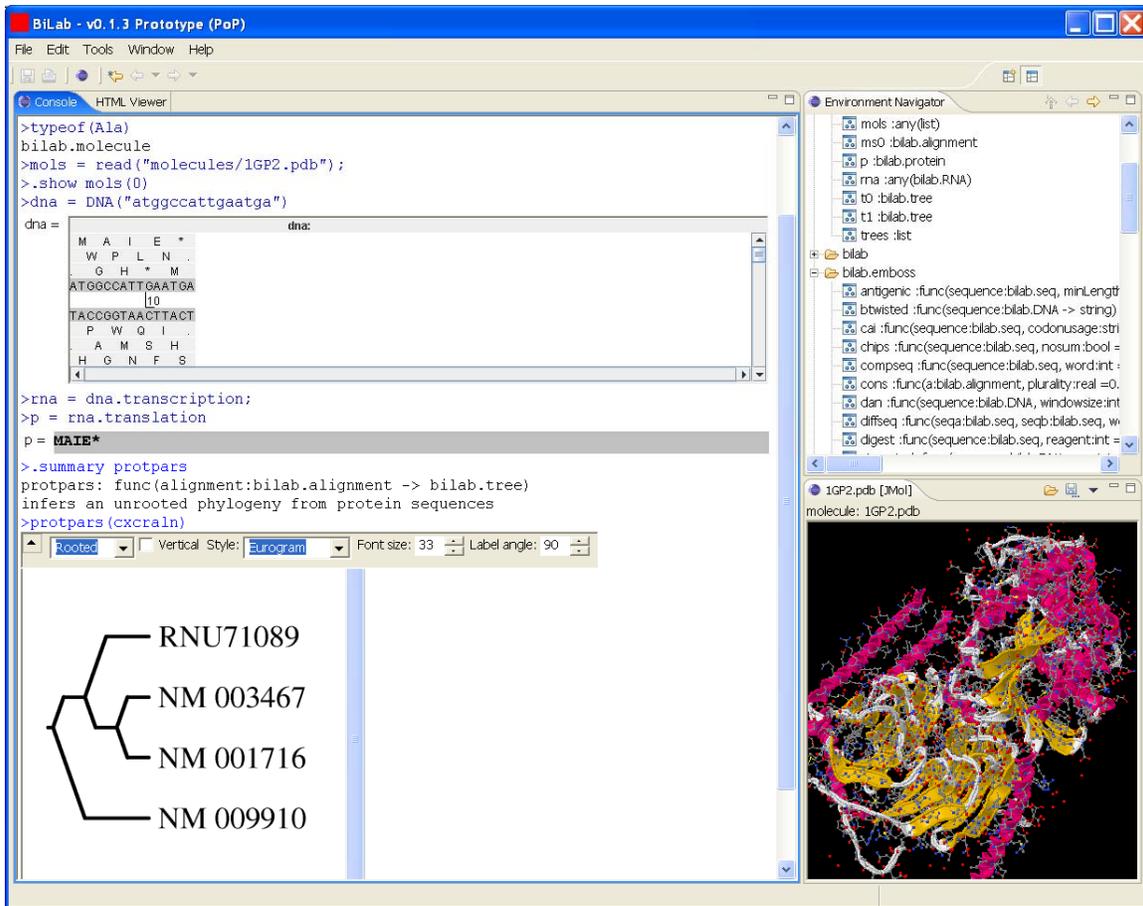


Figure 3: Screenshot of Proteomic Toolshop manipulating biological objects – DNA, RNA, proteins, trees, and molecules.

Protein Interaction Network Inference

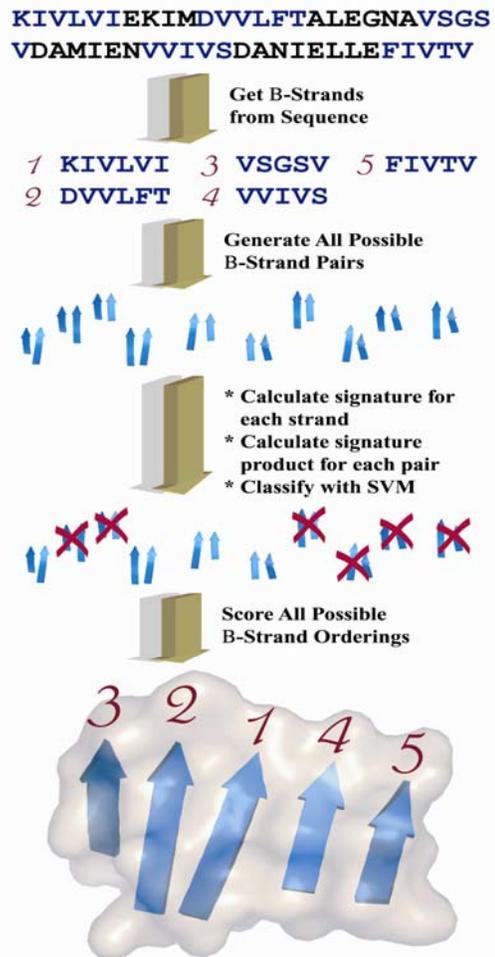
Sandia National Laboratories

Jean-Loup Faulon, Shawn Martin, W. Mike Brown

As part of Subproject 4, Systems Biology for *Synechococcus* Sp., in the Sandia National Laboratories-Oak Ridge National Laboratory Genomics:GTL project “Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling” we have developed computational tools to infer protein interaction networks. This work is relevant to *Synechococcus* Genomics:GTL (SGTL) because protein interactions are important for elucidating the functions of molecular machines and for understanding signaling pathways and gene regulation, all of which are critical to gain an understanding of the cell at the system level.

Over the course of this project we have developed top-down and bottom-up approaches. With the top-down approach, we probe structural and dynamical properties of known or inferred protein networks, while with the bottom-up approach we predict protein-protein interactions from experimental data, including phage display and two hybrids (such experiments are being performed within Subproject 1 for *Synechococcus*).

Last quarter we applied our signature support vector method for the prediction of β -sheet topology, requiring the consideration of long-range interactions between β -strands that are not necessarily consecutive in sequence. These interactions are difficult to predict using *ab initio* folding methods, and we envision our method will reduce the three-dimensional search space of folding calculations. Our method is shown in the adjacent figure and is based on the signature descriptor, which has previously been used to successfully predict protein-protein interactions. We have demonstrated how the signature descriptor can be used in a Support Vector Machine to predict whether or not two β -strands will pack adjacently within a protein. We have also shown how these predictions can be used to order β -strands within β -sheets. Using the entire PDB database with a randomly chosen test set, we have achieved 80.9 percent accuracy in packing prediction, 81.3 percent accuracy in β -strand ordering, and 88.3 percent accuracy in predicting β -sheet edge strands. This particular accomplishment is relevant to SGTL because predicting β -strand ordering can improve the results of *ab initio* protein folding methods such as Rosetta, a subject of work by Charlie Strauss in Subproject 2.



After our signature support vector paper appeared in *Bioinformatics*, we have received more than ten requests for our support vector code. We have also received requests for the signature calculation code from various biotech and pharmaceutical companies, including AstraZeneca, Genentech, MDL, Daylight, and IntelliChem. We have started a collaboration with Dr. S. Rasheed (University of Southern California) and Dr. Asa Ben-Hur (University of Washington) to elucidate protein complexes from MS pulldowns data.

Computational Inference of Biological Networks in Cyanobacteria Genomes

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

¹University of Georgia, ²University of Missouri at Columbia, ³University of California at Riverside, and ⁴University of California at San Diego

As part of Subproject 3, Computational Methods Towards The Genome-Scale Characterization of *Synechococcus* Sp. Regulatory Pathways, in the Sandia National Laboratories-Oak Ridge National Laboratory Genomics:GTL project “Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling,” we are working on:

- Prediction of *cis* regulatory elements at genome scale in cyanobacteria
- Prediction of operons and regulons at scale in cyanobacteria

- Mapping orthologous genes across microbial genomes
- Mapping biological pathways across microbial genomes
- Identification of functional modules through comparative analysis of microbial genomes
- Prediction of nitrogen assimilation networks in cyanobacteria
- Prediction of cross-talk networks between nitrogen assimilation and photosynthesis processes in cyanobacteria

This work is relevant to *Synechococcus* Genomics:GTL (SGTL) because the developmental work on computational capabilities for (a) *cis* regulatory element prediction, (b) operon prediction and (c) functional module prediction is essential to building a capability for computational inference of regulatory networks. Also, the developmental work on computational capabilities for inference and modeling of biological networks represents one of the key goals of SGTL. Computational prediction and characterization of regulatory networks in cyanobacteria represents one of the key scientific goals of SGTL.

Over the course of this project, we have developed a suite of computational prediction tools in support of comparative genome analysis and inference of biological networks, including P-MAP, JPOP, Excavator, CUBIC, PRIME, PROSPECT-PSPP and Promoco, all of which are downloadable from our website. We have also developed a computational pipeline for inference of biological networks for microbial organisms and predicted a number of biological networks and their cross-talk networks with other important biological processes in cyanobacteria using the computational capabilities we have developed. We have published or submitted 30+ full papers in journals, refereed conference proceedings, and book chapters about the work of this project (one of them won the best paper award at the *14th International Conference on Genome Informatics* in December 2004).

Significant recent results include:

- Developed prediction software JPOP for operon prediction
- Developed prediction software PROMOCO for accurate prediction of *cis* regulatory elements in microbial genomes
- Developed computational prediction capabilities for accurate orthologous gene mapping
- Developed a computational capability for functional module prediction in microbial organisms
- Predicted and characterized nitrogen assimilation network in cyanobacteria
- Predicted and characterized the cross-talk network between nitrogen assimilation and photosynthesis
- Characterized the evolution of phosphate degradation pathways

The impacts to the GTL world outside of this project include making our tools available to GTL researchers and the development of close collaborations with :

- Jim Tiedje of Michigan State (Shewanella)
- Penny Chisholm of MIT (Prochlorococcus)
- Heidi Sophia & Tjerk Straatsma of PNNL (*cis* regulatory element prediction and protein structure prediction)

In addition, we have developed close collaborations with leading researchers in other domains of microbial genomics and biology, including:

- Mike Adams of UGA, pyrococcus
- Mary Ann Moran of UGA, *Silicibacter*
- Barny Whitman of UGA, mathenococcus

- Jiong Yang, CWRU, text mining in support of biological network inference
- Hong Qian, University of Washington, modeling of network dynamics
- Zhirong Sun, Tsinghua University, modeling of network dynamics
- Eberhard Voit, Georgia Tech, network topology prediction
- Michael Zhang, Cold Spring Harbor Lab, cis regulatory binding site prediction

Other contributions include our computational tools, which are downloadable from our website and our organization of the GTL session at CSB2004 conference.

NMR Studies in RuBisCO

University of Michigan

Hashim Al-Hashimi, Shawn Stevens, Erik R. P. Zuiderweg, Grover Yip

As part of Subproject 1, Experimental Elucidation of Molecular Machines and Regulatory Networks in *Synechococcus* Sp., in the Sandia National Laboratories-Oak Ridge National Laboratory (ORNL) Genomics:GTL project “Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling,” we continue to develop and apply nuclear magnetic resonance (NMR) methodology to support the first goal of the project, which includes high-throughput structural and dynamic characterization of the protein-protein complexes that govern cellular function. Our efforts have focused on expression, purification, and stabilization of both the small and large subunits of RuBisCO (*rbcS* and *rbcL* respectively) for NMR studies. In addition, we have continued efforts towards optimizing relaxation pulse sequences to characterize protein dynamics by NMR spectroscopy. This work is included in Subproject 1 and is associated with the experimental work ongoing in A. Martino’s laboratory at Sandia as well as computational methodology developed in collaboration with A. Gorin of ORNL and in the future, D. Roe of Sandia, both of whom are associated with Subproject 2. Our combined goal is to characterize protein-protein interactions by coordinating the use of diverse techniques and methodology (i.e., phage display, NMR spectroscopy, and computational methods) to obtain information with a speed and resolution that could not be attained by any of these techniques separately.

Obtaining high-quality NMR spectra of the RuBisCO small subunit was proving difficult, so we decided to purify the large subunit of the enzyme for spectroscopy trials. While expression levels were very good, we were unable to optimize conditions to synthesize soluble protein. Thus, several attempts were made to refold the protein obtained from the inclusion bodies. The best results were obtained by refolding while the protein was bound to the Ni-NTA column. However, upon elution, the protein immediately aggregated and precipitated in an essentially quantitative fashion. Therefore, we concluded that the large subunit of RuBisCO cannot stand alone, especially at the concentrations required for NMR studies.

Efforts then returned to the RuBisCO small subunit where a matrix of buffer conditions was used in an attempt to stabilize the protein in monomer form (see table). These efforts are ongoing and previous results indicate that pH levels ≥ 7.0 are favored. To this end, we are testing a number of buffers in the range of pH 7.0-8.0 in combination with mono- and divalent salts, as well as some osmolytes (arginine + glutamine, and glycine), which often can enhance protein folding and stability without interfering with function (i.e., protein/ligand recognition). We have promising preliminary results for buffer conditions with higher pH (~7) (20 mM Mg²⁺, 5 mM 2-mercaptoethanol, 25 mM arginine and glutamine). We are currently working around these conditions with ¹⁵N labeled protein to see if we can further optimize the NMR spectra of *rbcS*.

| buffer | pH | [Na+] | [Mg2+] | Arg + Glu | glycine |
|-----------|-----|-------|--------|------------|---------|
| phosphate | 7.0 | 50 mM | 20 mM | - | - |
| | | 50 mM | - | 25 mM each | - |
| | | 50 mM | - | - | 1 M |
| | 7.5 | 50 mM | 20 mM | - | - |
| | | 50 mM | - | 25 mM each | - |
| | | 50 mM | - | - | 1 M |
| | 8.0 | 50 mM | 20 mM | - | - |
| | | 50 mM | - | 25 mM each | - |
| | | 50 mM | - | - | 1 M |
| Tris | 7.5 | 50 mM | 20 mM | - | - |
| | | 50 mM | - | 25 mM each | - |
| | | 50 mM | - | - | 1 M |
| | 8.0 | 50 mM | 20 mM | - | - |
| | | 50 mM | - | 25 mM each | - |
| | | 50 mM | - | - | 1 M |
| HEPES | 7.0 | 50 mM | 20 mM | - | - |
| | | 50 mM | - | 25 mM each | - |
| | | 50 mM | - | - | 1 M |
| | 7.5 | 50 mM | 20 mM | - | - |
| | | 50 mM | - | 25 mM each | - |
| | | 50 mM | - | - | 1 M |

Table 1: Buffer conditions under which NMR spectra of rbcS will be measured.

The Zuiderweg group continues to work on developing reliable relaxation experiments that can be used to identify flexible residues on protein surfaces involved in protein-protein interactions. The manuscript, “Duty-Cycle Heating Compensation in NMR Relaxation Experiments,” by Grover N.B. Yip and Erik R.P. Zuiderweg has been submitted, and addresses an underlying problem with current NMR relaxation measurement protocols in maintaining constant sample temperature throughout the execution of the relaxation series. The article investigates the sources of these problems and proposes a solution to the sample heating issue by including a compensation / saturation loop at the beginning of the pulse sequence. The method is verified with ¹⁵N spin relaxation measurements for human ubiquitin. The ability to measure NMR relaxation properties of macromolecules reliably is a prerequisite for experimental studies of molecular dynamics and molecular interaction.

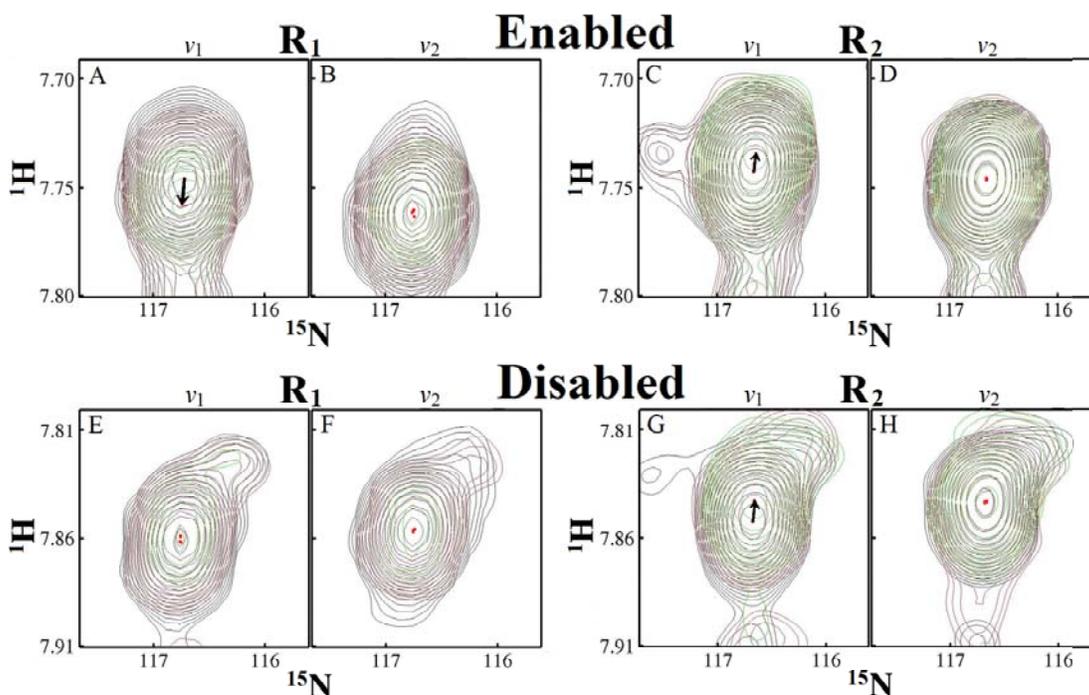


Figure 4: Each panel is an overlay contour plot of three spectra from a relaxation series for the ^1H - ^{15}N cross peak of residue 40 of human ubiquitin. The panels compare the results of *conventional* ($v1$) and our *improved* ($v2$) pulse sequences for T_1 and T_2 -CPMG with or without ^{15}N acquisition decoupling as indicated. Red dots illustrate the peak centers for the different spectra contours of varying relaxation times. Arrows indicate the shift direction as the relaxation time increases.

Our experimental efforts have primarily been impeded not so much by lack of NMR methods for characterizing protein-protein interactions, but by difficulties in preparing protein samples suitable for spectroscopic characterization by NMR. We believe this will be a general problem in studying proteins that interact with other proteins because many of these structures often fold upon recognition. There is therefore a tremendous need to carefully select protein targets that are amenable to both computational and experimental characterization.

The ability to characterize protein complexes by NMR is principally limited by the ability to prepare samples. This was not obvious at the onset especially since only recently have TROSY NMR methods been introduced which could allow characterization of such larger complexes. There is therefore a great need to develop computational methods that may utilize sequence information as a means of predicting the behavior of proteins under NMR sample conditions by providing information regarding the degree to which a given protein is expected to be structured.

Biopathways Graph Data Management System

Frank Olken (*Lawrence Berkeley National Laboratory*), **Nagiza Samatova** (*Oak Ridge National Laboratory*), **Michael Langston** (*University of Tennessee/ORNL*)

As part of Subproject 5, Computational Biology Work Environments and Infrastructure, in the Sandia National Laboratories-Oak Ridge National Laboratory Genomics:GTL project “Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling,” we are developing a graph data management system to support storage, analysis, and retrieval of biopathways data, metabolic, regulatory, and signaling networks. Within the *Synechococcus* Genomics:GTL (SGTL) project, biopathway networks are being inferred by Gorin (ORNL), Samatova (ORNL), and Xu (ORNL/University of Georgia). A biopathways graph data management (BGDM) system would be useful for inference, storage, and querying such datasets.

Current graph data management systems have very limited capabilities to answer graph optimization queries. We are working on graph query language design to support graph optimization queries. Such queries include shortest path queries, maximum weighted matchings, maximum cliques, and approximate graph matching queries. These queries involve the selection of a (sub)graph that minimizes or maximizes some objective function (e.g., cardinality, sum of edge weights [e.g., distances]) over a class of candidate (sub)graphs (paths, matchings, trees, etc.). Specifically, we have proposed a syntactical structure for such queries, and begun to consider issues of integration with classic selection queries and query optimization issues.

This work is being reported initially via a talk at the Department of Homeland Security (DHS) Conference: Working Together: R&D Partnerships in Homeland Security, April 27-28, 2005 in Boston.

Many Genomics:GTL projects, for example VIMSS, involve the identification and analysis of biopathways in microbial organisms. BGDM would be useful to these projects as well. Beyond Genomics:GTL, DHS and various intelligence and law enforcement agencies are interested in social network analyses (e.g., of communications traffic, surveillance) and for case-based retrieval of similar forensic cases.

Data Entry and Browsing (DEB) Tool

Lawrence Berkeley National Laboratory

Arie Shoshani, Vijaya Natarajan, Victor Havin

As part of the Sandia National Laboratories-Oak Ridge National Laboratory Genomics:GTL project “Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling,” the Scientific Data Management Group at Lawrence Berkeley National Laboratory (LBNL), in support of Subproject 5, Computational Biology Work Environments and Infrastructure, has developed a web-based Data Entry and Browsing (DEB) tool whose purpose is to facilitate capturing the metadata from experiments and laboratories and store them in a database in a computer searchable form. The key need is to have an easy-to-use intuitive system that integrates the metadata of all the related activities in this Genomics:GTL project. Specifically, the goal is to have a single system that captures the metadata for the Nucleotide Pool of microbes from the Scripps Institution of Oceanography (SIO), the microarray hybridization metadata from The Institute of Genomic Research (TIGR), and their analysis with the hyperspectral analyzer at Sandia.

The design of the DEB tool is based on input and insights from the biologists on the project. The interface design mimics the familiar laboratory notebook format. The system is built on top of the Object-Based Database Tools (the OPM database tools developed previously at LBNL) and the data is stored in the Oracle database system. Features of the DEB system that were developed prior to this report were described in the previous report in October 2004. This document reports progress since then.

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

New schemas were developed for the TIGR hybridization metadata (referred to as the TIGR schema) and Sandia's hyperspectral imaging metadata (referred to as the HSI schema). The TIGR hybridization schema that was developed in coordination with Kathy Kang from TIGR is modeled on the MIAME schema but it is much simpler because of our object-level modeling methodology. It consists of six object classes: Study, Probe_Source, Probe, Hybridization, Slide, and Scan, and the links between them. The HSI schema was developed based on input from Jeri Timlin, and has six object classes: HS_Experiment, HS_Slide, HS_Scan, Analysis_Entry, MCR_Analysis, and CLS_Analysis. Both schemas automatically generate relational databases in ORACLE, as well as the user-interfaces to data entry and browsing. For this purpose we used the DEB-generator tool we developed in the previous quarters.

Data entry for the TIGR hybridization metadata was achieved by using "dump files" from the TIGR database system. These files were generated by Kathy Kang and sent to LBNL staff. LBNL staff has developed a data loader that uses the provided files to populate the DEB database. During this quarter the number of entries loaded is: 13 studies, 178 hybridizations, 116 probe sources, 445 probes, 172 slides, and 173 scans. The data entry for the HSI database was done directly into the database by using the DEB user interface. So far, the number of entries completed by Jeri Timlin and Rachel Rhode (Sandia) is: 2 HS_Experiments, 10 HS_Slides, and 42 HS_Scans. In addition, new entries were entered for the SIO directly through DEB by Rob Herman and Chris Dupont.

The three databases for SIO (nucleotide pools), TIGR (hybridizations), and HSI (hyperspectral analysis) were developed independently of each other initially on the LBNL system. We migrated these databases to the SGTL facility in Oak Ridge National Laboratory (ORNL) (the "modpod" system), and then generated a combined DEB system for all three databases. The data from the three databases was loaded into the integrated DEB system, and linked. By "linking" we mean that the probes_source in the TIGR database "point" to the corresponding nucleotide_pools in the SIO database, and the HS_scans in the HSI database "point" to the scans in the TIGR databases. DEB facilitates such linking directly from the data entry interfaces.

The following new features were developed during this quarter:

- Generating new databases:
 - Automatic database builder – using command line
 - Can build/remove a database or just GUI
- New feature: inverse linking
 - Every forward link (data entry) can now be viewed as inverse

- New feature: “Go There”
 - Available for forward links – browse directly into the linked entry
 - Also available for inverse of links
- New features: Authorization
 - Previously – per object (instance) authorization
 - Added: *object-class* level authorization:
 - insert, and one of: read, read&update, read&update&delete
 - Instance level authorization was modified: Default – uses object class level authorization, Restricted – give any individual/group special permission
- Miscellaneous features
 - Password – now encrypted
 - Required* – advisory requirement for the existence of an entry
 - Choice of alphabetic ordering of entry lists

Our plans for the rest of 2005 include: 1) automatic path queries: given an object (e.g., HS_slide), find objects in another class (e.g., NP); 2) develop a report generator sufficient for use in publications; 3) develop a Schema generation tool, which will have a GUI that will guide the users on how to specify their schema as well as the UI layout. It will produce a specification in XML format, which will be used to drive the DEB database generator.

Event-driven, Hierarchical Modeling of *Synechococcus* Dynamics

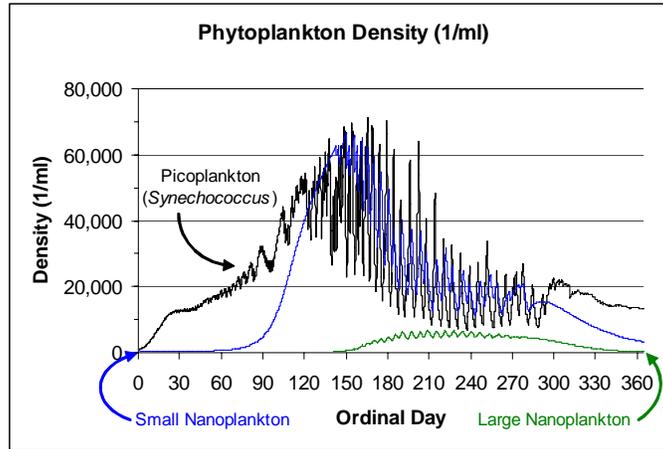
National Center for Genome Resources
 Damian Gessler, Andrea Belgrano

As part of Subproject 4, Systems Biology for *Synechococcus* Sp., of the Sandia National Laboratories-Oak Ridge National Laboratory Genomics:GTL project “Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling,” we are working on hierarchical modeling of carbon sequestration from oceanic carbon, nitrogen, and phosphorus concentrations to their role in carbon fixation in the carboxysomes of *Synechococcus*. We are also examining the effect of carbon sequestration on cell growth, population growth, and density dependent limitations on population growth and nutrient recycling back into the ocean.

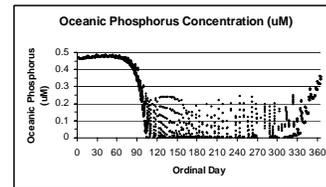
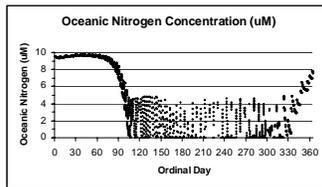
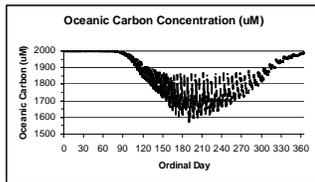
This work is relevant to *Synechococcus* Genomics: GTL (SGTL) because it connects intracellular process of carbon fixation being studied by other Subprojects to the larger, overall effect of picoplankton (*Synechococcus*) growth on marine carbon cycling.

To our knowledge, we are the first to combine nutrient stoichiometry and metabolic rate allometry in a single, unified dynamic model of marine carbon sequestration. Concurrently, we developed a new simulation approach for event-driven, disparate-time hierarchical modeling, implemented it into code, and used it to model the nutrient stoichiometry/metabolic rate allometry approach.

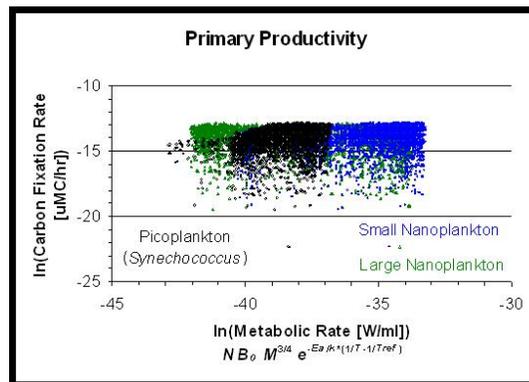
We have computational evidence, based on careful literature parameterization, for chaotic population dynamics of picoplankton (~0.9 μm ; *Synechococcus*) when modeling simultaneously with small (~6 μm) and large (~15 μm) nanoplankton in a single, competitive community. Chaotic dynamics are driven by seasonal temperature rises increasing picoplankton metabolic rates disproportionately relative to the other size classes. This eventually leads to a quasi-equilibrium as populations become limited by nitrogen and phosphorus.



As shown in the figure above, pico- and small nanoplankton show chaotic population dynamics, while large nanoplankton show only minor oscillations in carrying capacity. Large nanoplankton display a lag response and reach their population maximum after the summer peak in temperature and when the pico- and small nanoplankton populations are in decline. There is no explicit carrying capacity in the simulations; rather it is created as an emergent quasi-equilibrium between nutrient availability, light, and temperature driving cell growth and reproduction, and nutrient limitation, predation, and viral lysis driving cell mortality.



These three figures show that throughout the year, temperature changes smoothly from 10°C to 20°C, and then back to 10°C (not shown). Carbon is not limiting, but nitrogen and phosphorus limitation drive the chaotic dynamics observed in the population density graph.



Importantly, our simulations strongly support the Energy Equivalency Rule (above), which shows that as populations consume more carbon faster with increased temperature, the total carbon fixation rate for the community remains constant. This is because the *community* is fluxing carbon proportional to the product of population density and the average metabolic rate. Density is decreasing at the same proportion that metabolic rate increases.

Our full model is being reported elsewhere, and inspection shows that our hierarchical divide-and-conquer approach has significant benefits in modeling large, complex systems. All modular components of the model interact solely by supplying and consuming parameters, and firing and responding to events. Thus, for example, as our understanding of biotic and abiotic elements of carbon sequestration mature, we can incorporate them in a well-defined, encapsulated manner.

The Energy Equivalency Rule (Damuth 1981; Enquist, *et al.* 1998, Ernest, *et al.* 2003), if validated, is a powerful statement about community stability. If applicable to marine trophic levels, it may mean that we could understand and predict large-scale marine equilibria with only minor specifics about community assemblages, population dynamics, and cellular machines. This work increases our understanding of large-scale complex system modeling and the Energy Equivalency Rule regarding phytoplankton nutrient cycling.

Damuth, J. 1981. Population Density and Body Size in Mammals. *Nature* 290:699-700.

Enquist, B. J., J. H. Brown, and G. B. West. 1998. Allometric Scaling of Plant Energetics and Population Density. *Nature* 395:163-165.

Enquist, B. J., E. P. Economo, T. E. Huxman, A. P. Allen, D. D. Ignace, and J. F. Gillooly. 2003. Scaling Metabolism from Organism to Ecosystems. *Nature* 423:639-642.

Molecular Modeling of RuBisCO's Gating Mechanism

Sandia National Laboratories

Paul Crozier and Steve Plimpton

(In collaboration with Günter Wildner, Jürgen Schlitter, and Christian Burisch from Bochum University in Germany)

As part of Subproject 2, Computational Discovery and Functional Characterization of *Synechococcus* Sp. Molecular Machines in the Sandia National Laboratories-Oak Ridge National Laboratory Genomics:GTL project "Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling," we are working on targeted molecular dynamics (TMD) simulations of the gating of RuBisCO's binding pocket as a function of its sequence and its structure details. This work is relevant to the *Synechococcus* Genomics:GTL (SGTL) program because RuBisCO is at the heart of the carbon sequestration process and forms the bottleneck of photosynthesis. A better understanding of how RuBisCO gating rates control its specificity and reaction rate could lead toward enhancements of the carbon fixation process by molecular-level engineering of RuBisCO. Over the course of this project, we have developed the molecular simulation tools necessary for modeling RuBisCO's gating process and used those tools to predict gating free energy barriers for several varieties of RuBisCO, including *Synechococcus* mutants and WT.

We have recently completed the development of the TMD algorithm in our LAMMPS molecular dynamics software that has enabled the simulation of the gating mechanism for tobacco, rice, and *Synechococcus* WT RuBisCOs as well as the following mutant forms of *Synechococcus* RuBisCO: PDK, DK, ET470-471PA,

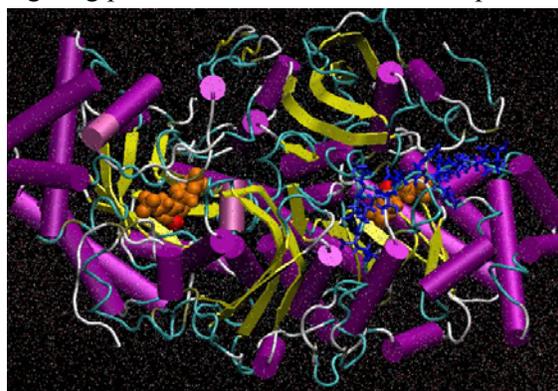


Figure 5: 2 RuBisCO large subunits with C-term tail in blue, sugars in orange, and Mg²⁺ coenzymes in red.

ETMDKL470 - 475PAMDTV, K474T, D473A. Tabulation of part of the electrostatics calculations within LAMMPS is an algorithm improvement that has enabled substantially faster MD simulation. These and other recent improvements to the LAMMPS software have facilitated RuBisCO gating simulations and led to significant recent results that show how small changes in the molecular structure of RuBisCO's C-terminal tail can dramatically affect the enzyme's observed operating characteristics. Tools developed as part of this project are freely downloadable and directly usable in computational molecular biophysics projects in other Genomics:GTL projects and in the entire molecular biophysics community. Improved understanding of RuBisCO's gating mechanism may lead to tailoring of the enzyme for better performance in native environments.

Manuscripts in preparation:

“Opening a Highly Secured Binding Pocket: Targeted Molecular Dynamics Studies of RuBisCO,” Christian Burisch, Paul S. Crozier, Günter F. Wildner, and Jürgen Schlitter.

“Macroecology and Macronutrients Limitation in Autotrophic Bacteria: A Perspective from Microbial Kinetics to Ecosystems Function,” Andrea Belgrano, Mick Follows, Paul Crozier, Damian Gessler, et al.

Spatial Modeling of *Synechococcus* Ecosystems

Sandia National Laboratories and National Center for Genome Resources

Shawn A. Means with Damian Gessler and Andrea Belgrano

As part of Subproject 4, Systems Biology for *Synechococcus* Sp., in the Sandia National Laboratories-Oak Ridge National Laboratory Genomics:GTL project “Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling,” we are working on implementing the nonspatial ecological models developed at the National Center for Genome Research (NCGR) in a reaction-diffusion model utilizing the reacting flows numerical solver MPSalsa. Spatial aspects are not included in the literature where extensive studies on nutrient limitations on population dynamics for *Synechococcus* – and other cyanobacteria. We therefore are currently involved in converting the rather detailed nonspatial model developed at NCGR into a corresponding model that includes the spatial distributions of these critical nutrients, such as nitrogen and phosphorus as well as carbon. As far as we are aware, such spatial modeling of these ecosystems has not been performed.

This work is relevant to *Synechococcus* Genomics:GTL (SGTL) because the impact of *Synechococcus* on a large ecological scale necessarily requires the study of large populations of this cyanobacteria and their interaction with their environment, hence our interest in studying an ecosystem with *Synechococcus*. As stated above, the spatial element and its impact on the population dynamics is relatively unknown and we intend to explore this aspect, which may hold implications for the impact of *Synechococcus* on the environment as a whole.

This work is in an initial phase. We are converting the model developed by the NCGR team into a form both understandable and appropriate for the spatial model. We are currently testing this conversion in a straightforward nonspatial manner for ensuring the proper implementation of the rather detailed aspects of this ecological model, which include distributions of nutrients, temperature, photon availability, and possibly fluid flows. We anticipate successful implementation of the spatial model during the subsequent quarter.

The impact to the Genomics:GTL world outside of this project is possibly the recognition of a need for spatial aspects in the large-scale ecological modeling of a small organism's impact on the biosphere and perhaps recognition of the impact of a small organism in regulating the biosphere at large, without necessary consideration of spatial elements.

Spatio-temporal Cell Modeling with Particles

Sandia National Laboratories

Steve Plimpton and Alex Slepoy

As part of Subproject 4, Systems Biology for *Synechococcus* Sp., of the Sandia National Laboratories-Oak Ridge National Laboratory Genomics:GTL project "Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling," we have been developing a simulator called ChemCell that tracks the spatial and temporal variations in concentrations of protein species as the molecules (represented as particles) diffuse within a cellular geometry and biochemically react with each other. This work is relevant to *Synechococcus* Genomics:GTL (SGTL) because we are developing a tool that could enable metabolic, signaling, or regulatory pathways of various kinds to be simulated in a realistic cellular environment for a variety of microbial cells.

Over the course of this project, we have developed an initial version of the tool which uses a heuristic Monte Carlo method for performing reactions each timestep between nearby reactants. We showed that in the limit of large diffusion rates (reaction-limited pathways), the heuristic method agrees with the nonspatial stochastic simulation algorithm (SSA) of Gillespie and its assumption of well-mixed reactants. We have also formulated the particle diffusion and reaction algorithms in ChemCell in a parallelizable form so that large problems can be run on parallel machines, and have demonstrated the code's scalability to a few hundred processors.

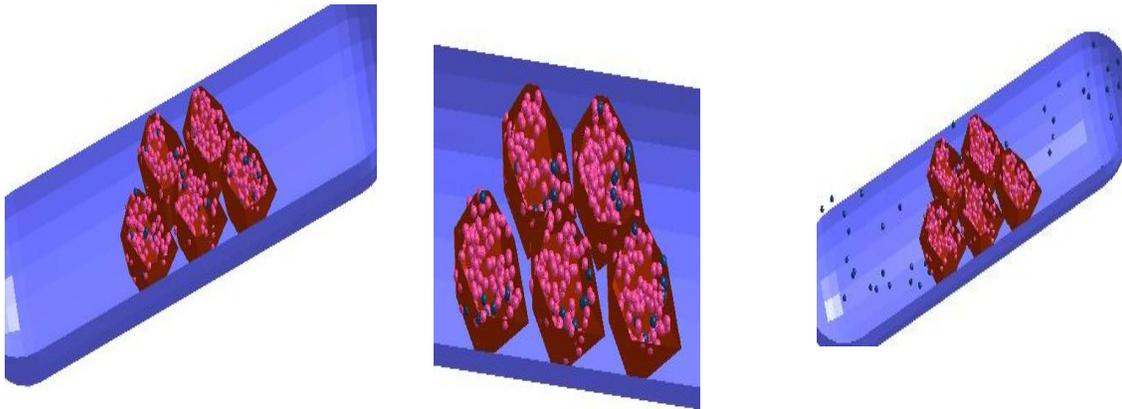
In the last few months we have worked in two areas. The first is to implement additional reaction algorithms within the ChemCell framework, the first of which is the nonspatial Gillespie SSA. This algorithm can now be run as an option within ChemCell and is useful for rapid prototyping of a reaction pathway and for comparison to a more detailed spatial simulation. The second is a new algorithm we have developed and now implemented that is a spatial form of the SSA. For each diffusive timestep, a Gillespie-style model is invoked to perform a sequence of reactions anywhere in the simulation domain, but with reaction probabilities computed and updated using neighbor information about nearby particles and their diffusion coefficients. This enables a more rigorous simulation of a reaction network without the hard-to-quantify errors that can potentially occur in ours (and other's) heuristic methods that use a finite timestep. We are currently testing the algorithm to compare its accuracy to other methods and to improve its computational efficiency. We view it as a useful alternative method for simulations to use within the ChemCell framework.

This quarter we also performed new simulations of the carbon-fixation pathway model for *Synechococcus* that we developed previously. In the new models, we changed the location of a key enzyme, carbonic anhydrase (CA), which converts bicarbonate ions to carbon dioxide. It is unknown whether CA exists only in the carboxysome as an effective carbon concentration mechanism or also in the cytoplasm. We also experimented with layering the CA and RuBisCO enzymes within the carboxysomes, as some experimental data indicates the CA may exist preferentially at the carboxysome inner surface. Images of the various ChemCell models are shown in the accompanying figure. We found that glucose production (the end product of the

pathway being modeled) was reduced by a factor of 2-times when CA is only present in the cytoplasm versus the carboxysome. And glucose production was increased by 30-50 percent when the CA was layered at the carboxysome surface. While these results are qualitative, they indicate how ChemCell can be used to include spatial realism in a pathway model.

We have applied to DOE for permission to release ChemCell as an open source code. We will be working this summer to polish a releasable version with additional tools for setting up simulations and analyzing their results. Our goal, by the end of this fiscal year, is to provide a tool for Genomics:GTL and other cell biology researchers that can be used and extended for a variety of interesting modeling problems where spatial and temporal resolution is needed.

Figure 6: Three snapshots from ChemCell models with different carbonic anhydrase (CA) representations. On the left, the CA is mixed randomly in the carboxysomes with RuBisCO enzymes; in the middle it is layered against the inner surface of the carboxysomes; on the right it is present in the cytoplasm of the *Synechococcus* cell.



Publications

- Belgrano, A., U. M. Scharler, J. Dunne, R. E. Ulanowicz, “Aquatic Food Webs: An Ecosystem Approach,” *Oxford University Press*, Oxford, 2005.
- Chen, X., Su, Z., Xu, Y. Jiang, T., “Computational Prediction of Operons in *Synechococcus* sp WH8102,” *Proceedings of the 14th International Conference on Genome Informatics*, pp. 211 – 222, best paper award, 2004.
- Chen, X., Su, Z., Dam, P., Palenik, B., Xu, Y., Jiang, T., “Operon Prediction by Comparative Genomics: an Application to the *Synechococcus* WH8102 Genome,” *Nucleic Acids Res*, Vol 32 (7), 2147 – 2157, 2004.
- Chen, Z., Xu, Ying, “Multi-Scale Hierarchical Structure Prediction of Helical Transmembrane Proteins,” *Proceedings of IEEE Computational Systems Bioinformatics Conference*, (in press), 2005.
- Dam, P., Su, Z., Olman, V., Xu, Y., “In silico Construction of the Carbon Fixation Pathway in *Synechococcus* sp. WH8102,” *Journal of Biological Systems* (invited paper), Vol 12, pp 97 – 125, 2004.
- Guo, J.T., Ellrott, K., Chung, W., Xu, D., Passovets, D., Xu, Y., “PROSPECT-PSPP: An Automatic Computational Pipeline for Protein Structure Prediction,” *Nucleic Acids Research*, Vol 32, W1 – W4, 2004.
- Huang, J., Xu, Y., Gogarten, J.P., “Ancient Lateral Gene Transfer Marks the Opisthokont,” submitted, *Molecular Biology and Evolution*, 2005.
- Huang, J., Su, Z., Xu, Y., “The Evolution of the Phosphonate Degradation Pathways,” *Journal of Molecular Evolution*, (in press) 2005.
- Li, G., Liu, Z., Olman, V., Xu, Y., “A Fast and Accurate Algorithm for Prediction of *cis* Regulatory Elements,” submitted, *Nucleic Acids Research*, 2005.
- Li, G., Lu, J., Olman, V., Xu, Y., “Highly Sensitive Algorithm for Identification of *cis* Regulatory Elements: from Cliques to High Information Content Analysis,” submitted, *Bioinformatics*, 2005.
- Li, G., Wang, X., Qi, X., Zhu, B., Xu, Y., “A Linear-time Algorithm for Computing the Translocation Distance Between Signed Genomes,” *Theoretical Computer Science*, (in press) 2005.
- Liu, Z., Chen, D., Bensmail, H., Xu, Y., “Gene Expression Data Clustering with Kernel Principal Component Analysis,” *Journal of Bioinformatics and Computational Biology*, Vol 3(2), 303-316, 2005.
- Liu, Z., Chen, D., Xu, Y., “Logistic Support Vector Machines and Their Application to Gene Expression Data,” *Journal of Bioinformatics Research and its Applications*, (in press), 2005.
- Liu, Z., Mao, F., Guo, J, Yan, B., Xu, Y., “Quantitative Validation of Protein-DNA Interaction in Transcription Process: Distance-dependent Knowledge-based Potential Considering Multi-body Interaction,” submitted, *JBC*, 2004.
- Liu, Z., Mao, F., Guo, J., Yan, B., Xu, Y., “Quantitative Validation of Protein-DNA Interaction in Transcription Process: Distance-dependent Knowledge-based Potential Considering Multi-body Interaction,” *Nucleic Acids Research*, 32:2, pp 546-558, 2005.

- Mao, F., Wu, H., Olman, V., Xu, Y., “Accurate Prediction of Orthologous Gene Groups in Microbes,” *Proceedings of IEEE Computational Systems Bioinformatics Conference*, (in press) 2005.
- Mao, F., Su, Z., Olman, V., Dam, P., Liu, Z., Xu, Y., “Mapping of Orthologous Genes in the Context of Biological Pathways: an Application of Integer Programming,” submitted, *PNAS*, 2005.
- Martin, S., Roe, D., Faulon, J.-L., “Predicting Protein-Protein Interactions Using Signature Products,” *Bioinformatics*, 21:218-226, 2005.
- Olman, V., Peng, H., Su, Z., Xu, Y., “Mapping of Microbial Pathways Through Constrained Mapping of Orthologous,” *Proceedings of The IEEE Computational Systems Bioinformatics Conference*, pp 363 – 370, 2004.
- Park, B.H., Dam, P., Pan, C., Xu, Y., Geist, A., Heffelfinger, G., Samatova, N., “*In silico* Recognition of Protein-protein Interactions: Theory and Applications, Book Chapter in Advanced Data Mining Technologies in Bioinformatics,” *Idea Group Publishing*, (in press) 2005.
- Song, Y., Liu, C., Huang, X., Malmberg, R., Xu, Y., Cai, L., “Efficient Parameterized Algorithm for Biopolymer Structural-Sequence Alignment,” submitted, *WABI*, 2005.
- Su, Z., Olman, V., Mao, F., Xu, Y., “Comparative Genomics Analyses of ntcA Regulons in Cyanobacteria: Regulation of Nitrogen Assimilation and its Coupling to Photosynthesis,” submitted, *Nucleic Acids Research*, 2005.
- Su, Z., Dam, P., Mao, F., Olman, V., Palenik, B., Paulsen, I., Xu, Y., “Computational Inference and Experimental Validation of Nitrogen Assimilation Regulatory Networks in Cyanobacterium *Synechococcus sp.* WH8102,” submitted, 2005.
- Su, Z., Dam, P., Mao, F., Chen, X., Olman, V., Jiang, T., Palenik, B., Xu, Y., “Towards Computational Inference of Regulatory Pathways in Prokaryotes: An Application to Phosphorus Assimilation Pathways in *Synechococcus sp.* WH8102,” submitted, *Genomes Research*, 2004.
- Wu, H., Su, Z., Olman, V., Xu, Y., “Prediction of Functional Modules Through Comparative Genome Analysis and Application of Gene Ontology,” *Nucleic Acids Research*, (in press) 2005.
- Yan, B., Pan, C., Olman, V., Hettich, B., Xu, Y., “A Graph-theoretic Approach to Separation of b- and y-ions in Tandem Mass Spectra,” *Bioinformatics*, 21: 563-574, 2005.
- Yan, B., Qu, Y., Mao, F., Olman, V., Xu, Y., “PRIME: A Mass Spectrum Data Mining Tool for *De Novo* Sequencing and PTMs Identification,” *Journal of Computer Science and Technology* (in press) (by invitation) 2005.
- Yan, B., Pan, C., Hettich, B., Olman, V., Xu, Y., “Separation of Ion Types in Tandem Mass Spectrometry Data Interpretation –a Graph-Theoretic Approach,” *Proceedings of The IEEE Computational Systems Bioinformatics Conference*, pp. 236 – 244, 2004.
- Yip, G. N., Zuiderweg, E.R., “Duty-Cycle Heating Compensation in NMR Relaxation Experiments,” submitted.
- Yip, G. N., Zuiderweg, E.R., “A Phase Cycle Scheme that Significantly Suppresses Offset-Dependent Artifacts in the R2-CPMG 15N Relaxation Experiment,” *J Magn Reson.* (1):25-36. 2004.

Presentations

- Belgrano, A., "Ecological Modeling," *GTL Semi-annual meeting*, Seattle, WA, 2005.
- Faulon, J.-L., Brown, W.M., Martin, S., "Biological and Chemical Structures Inference and Design," *Biocomputing at UNM*, Albuquerque, NM., April 22, 2005.
- Faulon, J.-L., Brown, W.M., Martin, S., "Inverse Problems in Cheminformatics and Bioinformatics," *SIAM SCE05*, Orlando, FL, February 12-15, 2005.
- Faulon, J.-L., Brown, W.M., Martin, S., "Reverse Engineering Chemical Structures from Molecular Descriptors: How Many Solutions?," *ACS National Meeting*, San Diego, CA, March 13-17, 2005.
- Gessler, D., "Event-driven, Hierarchical Modeling of *Synechococcus* Dynamics," *GTL Semi-annual meeting*, Seattle, WA, 2005.
- Martin, S., Davidson, G., May, E., Werner-Washburne, M., Faulon, J.-L., "Inferring Genetic Networks from Microarray Data," *New Mexico Bioinformatics Symposium*, Santa Fe, NM, March 31-April 1, 2005.
- Martin, S., Brown, W.M., Strauss, C., Rintoul, M.D., Faulon, J.-L., "Predicting Protein-Protein Interactions Using Signature Products with an Application to β -Strand Ordering," *Genomes to Life Contractor-Grantee Workshop III*, Washington, D.C., February 6-9, 2005.
- Plimpton, S. J., Slepoy, A., *GTL Program Meeting*, Washington DC, February, 2005.
- Plimpton, S. J., Slepoy, A., *SIAM Computational Science & Engineering Conference*, Orlando, FL., February, 2005.
- Plimpton, S. J., Slepoy, A., *Symposium on Computational Cell Biology Conference*, Lenox, MA., March, 2005.
- Xu, Y., "Computational Prediction of Biological Structures and Functions: From Protein Complexes to Biological Networks," *Winship Cancer Center*, Emory University (invited by Director and Prof. Jonathan Simon), 2005.
- Xu, Y., "Computational Prediction of Functional Units and Regulatory Networks in *Cyanobacteri*," Invited talk, *Aquatic Sciences Meeting (ASLO)*, Salt Lake City, UT., 2005.
- Xu, Y., "Inference of Complex Regulatory Networks in Microbes Through Comparative Genome Analysis and Microarray Data Analysis," *Biological Science Department, University of Southern California* (invited by Prof. Fengzhu Sun), 2005.
- Xu, Y., "Predicting and Modeling Microbial Regulatory Networks Using Dynamical Models," *Departmental Seminar, EECS Department, Case Western Reserve University* (invited by Prof. Jiong Yang), 2005.

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