SAND REPORT
SAND2004-0038
Unlimited Release
Printed January 2004

Genomes to Life Project Quarterly Report October 2003

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.

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

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

SAND2004-0038
Unlimited Release
Printed January 2004

Genomes to Life Project Quarterly Report October 2003

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

Abstract

This SAND report provides the technical progress through October 2003 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], Hashimi M. Al-Hashimi[8], Paul Crozier[1], George S. Davidson[1], Jean Loup Faulon[2], Laurie J. Frink[1], Damien Gessler[12], David M. Haaland[1], Bob Harrington[2], William E. Hart[1], Erik Jakobsson[7], Todd Lane[2], Tao Jiang[9], Shawn Martin[1], Frank Olken[4], Brian Palenik[6], Hoony Park[3], Steven J. Plimpton[1], Diana C. Roe[2], Nagiza F. Samatova[3], Arie Shoshani[4], Charlie E. M. Strauss[5], Peter Steadman[12], Edward V.Thomas[1], Jerilyn A. Timlin[1], 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

Oak Ridge National Laboratories, Sandia National Laboratories, Lawrence Berkeley National Laboratories, Los Alamos National Laboratories, University of California San Diego (Scripps), University of California Riverside, University of Michigan, University of Illinois Urbana/Champaign, National Center for Genome Resources, Molecular Sciences Institute, Joint Institute for Computational Sciences, University of Missouri.

Executive Summary

Technical Progress

Our experimental efforts continue to be focused on developing experimental protocols for *Synechococcus*. To this end, we continue to optimize the separation process for ultrasmall *Synechococcus* inclusion bodies such as the carboxysome. Other efforts are focused on developing protocols for protein-protein interactions within the *Synechococcus* carboxysome via bacterial 2-hybrid methods, gene tags for pull down experiments, and phage display. We also carried out hyperspectral scanning of 250 gene *Synechococcus* microarrays and interacted with TIGR to improve the hybridization process and to supply them with the *Synechococcus* WH8102 DNA for the full genome microarray that they will be fabricating for us.

Our molecular machines computational biology effort made significant progress this quarter on the development of computational analysis tools for mass spectrometry data, demonstrating reliable ion identification in raw MS spectra using computational methods for the first time. We developed a novel algorithm, `KeyGeneMiner`, for the identification of key genes responsible for the oxygenic photosynthetic process in cyanobacterial genomes, including *Synechococcus*. We also improved the Rosetta program to treat long loops, and implemented a new optimization algorithm for our `LAMMPS` molecular dynamics code to enable more rapid molecular structure equilibration. We enhanced our docking code, `PDOCK`, by integrating it with the `Coliny` optimization toolkit to access a wide range of global optimizers to enable modeling of peptides with both continuous backbone torsion angles and discrete choices of side chains from a rotomer library.

Progress in our effort to develop new and effective protocols for systematic characterization of regulatory pathways includes the development of a preliminary version of a computational pipeline for interpretation of multiple types of biological data for biological pathway inference. We have also prototyped these methods on *Synechococcus* WH8102 to make several predictions, including a signaling/regulatory network for the phosphorus assimilation pathway and a genome-scale protein-protein interaction map.

Our computational systems-biology effort has produced a prototype simulation framework for a discrete particle representation of a *Synechococcus* cell. By incorporating many of the intra-cellular chemical and physical phenomena relevant to *Synechococcus*'s carbon fixation behavior in this model we have produced a model which yields not only a quantitative sense of how reaction rates and concentrations affect the simulation, but also a more fundamental physical understanding of the underlying processes and how they are affected by the geometry of the cell (see http://www.genomes-to-life.org/highlights for simulation movies). Other systems biology highlights include the discovery of a rigorous mathematical method for making predictions regarding protein-protein interactions across related organisms using sequence data.

Our computational work environments and infrastructure effort established a data and computational infrastructure for use by the project with the installation of several pieces of computing hardware at ORNL. We also initiated development of tools for browsing protein interaction networks and identifying functional association links between proteins and designed a language for our Matlab-like bioinformatics work environment ("BiLab").

Project Management and Progress Toward Milestones

Our large group met for two days in Berkeley, CA, October 21 and 22, 2003. Our next meeting is scheduled in March 2004 and will be held in Phoenix, Arizona. We have continued our monthly executive team teleconferences to discuss project management issues and our monthly technical team teleconferences to discuss technical progress on the project. Progress toward project milestones tracking with the proposal fairly well, although we continue our efforts to catch up and recover from the FY02 budget problems.

Experimental Elucidation of Molecular Machines and Regulatory Networks in *Synechococcus* Sp. (Subproject 1) Highlights

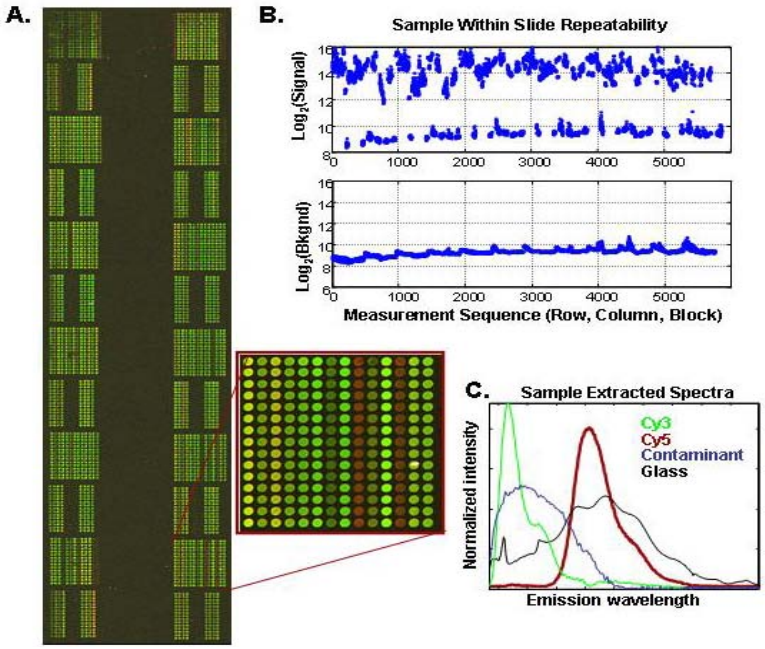


Fig. 1-1. 250-gene *Synechococcus* microarray data analysis. A. Cy5/Cy3 image (commercial microarray scanner, inset is single block enlarged). B. Statistical analysis of within slide repeatability. C. Sample of pure component spectra extracted from a hyperspectral image of microarray with green contaminant present.

This quarter we continued our efforts to optimize the separation of the ultrasmall *Synechococcus* inclusion bodies such as the carboxysome, and current purifications show (by Western blots) enrichment for RuBisCO. We have also begun developing and applying experimental protocols to determine the protein-protein interactions within the *Synechococcus* carboxysome including bacterial 2-hybrid methods, gene modifications to tag cso operon genes for pull down experiments, and phage display. Initial bacterial 2-hybrid results for 10% of the 10x10 matrix of potential interactions have not indicated positive interactions. A targeted cso operon phage library (7-mers and 12-mers) is being constructed for phage display experiments, three TPR repeat units were cloned from PSA E proteins and will be synthesized as targets. In addition, we carried out statistical analysis and hyperspectral scanning of 250 gene *Synechococcus* microarrays, and interacted with TIGR to improve the hybridization process and supplied them with the *Synechococcus* WH8102 DNA for the full genome microarray that they are fabricating for us. Construction has begun on our new optimized hyperspectral microarray scanner and progress has been made toward the use of multiple green fluorophores in microarrays.

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).

Oak Ridge National Laboratories, Sandia National Laboratories, Lawrence Berkeley National Laboratories, Los Alamos National Laboratories, University of California San Diego (Scripps), University of California Riverside, University of Michigan, University of Illinois Urbana/Champaign, National Center for Genome Resources, Molecular Sciences Institute, Joint Institute for Computational Sciences, University of Missouri.

Accomplishments

Carboxysome Isolation and Purification: The carboxysome preparation was modified to optimize separation of the ultrasmall *Synechococcus* inclusion bodies (75 nm in diameter as determined by cryogenic tomography). The protocol is highlighted by the use of two high-speed centrifugation steps separated by addition of the divalent $MgSO_4$ ion. The high speeds in combination with Mg^{2+} ions collapses the electric double layer around the carboxysome leading to floculation of the inclusion bodies. The particulate matter can be separated and further purified. Purification protocols are currently being optimized. Analysis of the soluble and particulate fractions of the protein lysates shows the enrichment of RuBisCO by blotting with a RuBisCO large antibody indicating the enrichment for the carboxysome. This quarter, we have designed and ordered antibodies to other proteins coded by the cso operon. These proteins will help further characterize carboxysome enrichment.

Interactions within the Carboxysome: Experiments to determine interactions are occurring on three fronts. Initial bacterial 2-hybrid results have not indicated positive interactions. Eight different genes in the cso operon have been cloned into bait and target vectors of the Stratagene bacterial 2-hybrid systems. Thus far, rbcL has been tested against rbcL and rbcS. Also, csoS2 has been tested against ccmk1, rbcL, rbcS, csoS2, cso S3, and peptide A. These are only initial results; sequencing, expression, and condition testing has not been completed.

Genetic modifications to tag cso operon genes for pull down experiments continue. Primers have been designed to clone the operon genes into Invitrogen's TOPO isomerase directional vectors. His tags have been built into the vectors. Also, work has begun to build multiple cloning sites into pMUT100 and pRL153. pMUT100 and pRL153 are suicide and replicative vectors that have been used in conjugation with transformations of *Synechococcus*. They have limited cloning digest sites and introduction of a MCS to the vectors will assist in later cloning efforts.

Finally, a targeted cso operon phage library is being constructed for phage display experiments. The cso operon was cloned into Statagene's pBluScript. The cso operon has been digested out of pBS, gel isolated, and sonicated to form 200-400 bp random oligonucleotides. Next quarter, the randomized cso fragments will be ligated into two different phagemids. Also, RuBisCO large and small subunits have been cloned and are in the process of being synthesized as target proteins phage display experiments.

Protein Binding Domains and Interaction Networks: Initial control phage display experiments with 7-mer and 12-mer libraries were completed. Three TPR repeat units were cloned from PSA E proteins, and the TPR peptides are in the process of being synthesized as targets for phage display experiments.

Gene Regulatory Network by Microarray Analysis: We have worked on the construction of a mutant in an orf that might be a phosphonatase (degrades C-P bonds), have supplied WH8102 DNA for the full genome microarray construction, and have performed statistical analysis and hyperspectral scanning of 250 gene *Synechococcus* microarrays.

We have been working on the construction of a mutant in an orf that might be a phosphonatase. This orf was predicted to be a potential phosphonatase by Ying Xu's group (ORNL and U Georgia) working in Subproject 3, based on `Prospect` analyses. We amplified the orf, ligated it into the vector pMut100 and conjugated it into WH8102. We have successfully obtained colonies in which the orf is presumably inactivated. We are now growing cells through a few transfers in regular media before testing for phosphonatase activity. The mutants will be used in microarray gene regulatory experiments. We have provided RNA for wild-type cells and two mutants in which the phosphate regulatory system

(PhoB/PhoR) has been inactivated. Again this will help tie-in with phosphate pathway modeling by Ying Xu's group. These will presumably be some of the first RNAs used on the whole genome microarray.

In this quarter we also performed detailed statistical analysis of some of the 250 gene TIGR *Synechococcus* microarrays (including dye-flip experiments) and have noted anomalies (high background, specks, block-to-block variation, etc). This work has led to the development of tools for visualization of statistical properties of microarray slides. We have made recommendations to TIGR to improve their microarray hybridization process based on our graphical statistical analysis of these data and will continue this iterative process in the next quarter as the full genome arrays are printed. Follow-up hyperspectral scans have been performed in order to determine the source of the irregularities; multivariate data analysis will be completed in the next quarter. We have also scanned several slides from other GTL laboratories (ORNL, PNL, Michigan State) to explore their quality and will finish analysis of these slides in future quarters and have begun the needed work to increase the throughput of microarrays through the use of multiple green dyes and hyperspectral scanning. Also during this quarter, we purchased the equipment to build the new, optimized hyperspectral scanner. The majority of the equipment has arrived and construction has begun. Finally, we continued making improvements in our data acquisition and multivariate analysis software and have purchased an upgrade to the microarray analysis software.

NMR Analysis of Protein-Protein Interactions: Oak Ridge National Laboratory and University of Michigan are developing and applying solution NMR methods based on measurements of residual dipolar couplings (RDCs) for high throughput structural and dynamic characterization of protein-protein interactions. These methods are expected to accelerate NMR characterization of protein complexes through the integration of computational biology during protein structure analysis. The long-term goals of this work are to investigate protein-protein interactions relevant to the protein RuBisCO from *Synechococcus* and carboxysome. Novel NMR methods and attendant computational data analysis tools are also being developed to probe interactions in these proteins using peptide moieties that are attached to field oriented phage particles as a structural complement to a phage display. The latter effort is in collaboration with experimental work conducted at Sandia National Laboratory (Subproject 1) and computational methodology at Oak Ridge National Laboratory (Subproject 2).

Progress has been made in expressing and purifying the protein "small RuBisCO" which will be the initial target for application of the developed NMR methods. The `CAP` program (combinatorial assignment procedure) developed by ORNL group for high-throughput resonance assignments in protein systems using residual dipolar couplings has been tested successfully using "simulated" data and is now ready for implementation on the RDCs that will be measured on the protein RuBisCO.

We have also developed improved CPMG NMR pulse sequences used to measure conformational exchange dynamics processes at the multi-micro second time scale. These experiments and their improvements are pivotal to our proposed work with respect to identifying protein-protein binding interfaces on the basis of pre-existent induced-fit dynamic equilibrium. The Zuiderweg group at U Michigan (Subproject 1) has designed new NMR pulse sequence elements for this experiment, and by doing so increased its temperature stability, reliability and reproducibly. We are in the process of writing a manuscript on this subject.

Progress Towards Milestones

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

Aim 2. Tag central proteins of carboxysome (1/03). Genetic modifications to tag cso operon genes for pull down experiments continue. Primers have been designed to clone the operon genes into Invitrogen's TOPO isomerase directional vectors. His tags have been built into the vectors.

Aim 1. PCR amplify genes to be used as receptors in phage display. Design phage libraries. Begin testing (2/03). We are at various stages. Some of the genes have been PCR amplified and cloned (rbcL, rbcS, and 3 TPR repeats). Two phage libraries are complete and a third is in progress. Tests of positive controls are complete.

Aim 2. MS characterization of protein complexes. Determine consensus ligands (5/03). We have a good part of the carboxysome purification done and will be able to test by MS soon.

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). The new optimized hyperspectral scanner, delayed until capital equipment money was available in July 2003, has had all the optical components designed, all parts ordered, and the construction begun. In addition, four-green fluorophore microarrays have been generated, hybridized, and scanned with the hyperspectral scanner in the first step to improving the throughput of microarrays. Initial multivariate curve resolution of the 4-color slides demonstrates the ability to separate and quantitate these highly overlapped dyes.

Aim 2. Expression and purification of 15N-, and 15N/13C-, and 15N/13C/2H – isotopically enriched proteins (8/03). Progress continues towards purifying the protein RuBisCO. An expression plasmid has been developed and preparation of $^{15}N/^{13}C$ labeled RuBisCO is currently in the works.

Aim 2. NMR sample conditioning and optimization for free proteins and protein-protein complexes with and without liquid crystalline media (8/03). Five liquid crystalline media (purple membrane, phage, bicelles, polyacrylamide gel, and cellulose) have been prepared and tested for solvent compatibility (pH, temperature and salt concentration).

Collaboration With Others

There has been continued collaboration with Dean Price and Murray Badger of the Australian National University, and continued collaboration with Grant Jensen at Cal Tech to carryout electron microscopy and cryoelectron tomography studies of the carboxysome. This quarter we have also initiated 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 continue strong interaction with GTL team members. We continue to work with Arie Shoshani, LBNL (Subproject 5) in developing a laboratory data management platform. We have worked closely with the University of Michigan to develop experimental plans. Also, we have worked with Steve Plimpton, SNL (Subproject 2) and his team to discuss future docking calculations.

There has been continued collaboration with the Werner-Washburne and the Lyons groups at the University of New Mexico Biology Department and Health Sciences, respectively, to test our hyperspectral microarray scanner with their yeast and mouse genome arrays. Brian Palenik, UCSD/Scripps (Subproject 1), a member of our team and Ian Paulsen, TIGR (under contract with our

project) have continued to provide data from their 250 gene *Synechococcus* microarrays for hyperspectral scanning and statistical analysis. Several new collaborations to perform hyperspectral imaging and analysis for quality control have been established though Eugene Kolker, BiaTech, with other GTL laboratories, ORNL, PNL, and Michigan State using microarrays.

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 planned set of biophysical characterization tools for DOE Facility III.

Publications and Presentations

1. Gorin, Andrey and Al-Hashimi, Hashim M. "Computational Challenges in Rapid Characterization of Protein-protein Interactions by NMR Based Methods," *Proceedings of the International Moscow Conference on Computational Biology, Moscow State University* pp 86-87, 2003.

2. Haaland, David M., Keenan, Michael R., Van Benthem, Mark H., Timlin, Jerilyn A., Sinclair, Michael B., Kotula, Paul G., and Koehler, Frederick W. "Multivariate Curve Resolution applied to Hyperspectral Images," *Gordon Research Conference on Statistics in Chemistry and Chemical Engineering*, Mount Holyoke College, South Hadley, MA, Invited, July 27-August 1, 2003.

3. Haaland, David M., Timlin, Jerilyn A., Sinclair, Michael B., Van Benthem, Mark H., Thomas, Edward V., Martinez, M. Juanita, Aragon, Anthony D., and Werner-Washburne, Margaret, "Improving Microarray Analysis with Hyperspectral Microarray Scanning, Experimental Design, and Multivariate Data Analysis," *United States Army Research Institute for Infectious Diseases (USAMRIID)*, Fort Detrick, MD, Invited, May 7, 2003.

4. Haaland, David M., Timlin, Jerilyn A., Sinclair, Michael B., Van Benthem, Mark H., Martinez, M. Juanita, Aragon, Anthony D., and Werner-Washburne, Margaret, "Multivariate Curve Resolution for Hyperspectral Image Analysis: Applications to Microarray Technology," *Proceedings of SPIE Vol. 4959, Spectral Imaging: Instrumentation, Applications, and Analysis II*, pp 55-66, Edited by Richard M. Levenson, Gregory H. Bearman, Anita Mahadevan-Jansen, SPIE, Bellingham, WA, 2003.

5. Sinclair, Michael B., Timlin, Jerilyn A., Haaland, David M., Werner-Washburne, M. "Design, Construction, Characterization, and Application of Hyperspectral Microarray Scanner," *Applied Optics: Optical Technology and Biomedical Optics*, Submitted July 2003.

Computational Discovery and Functional Characterization of *Synechococcus* Sp. Molecular Machines (Subproject 2) Highlights

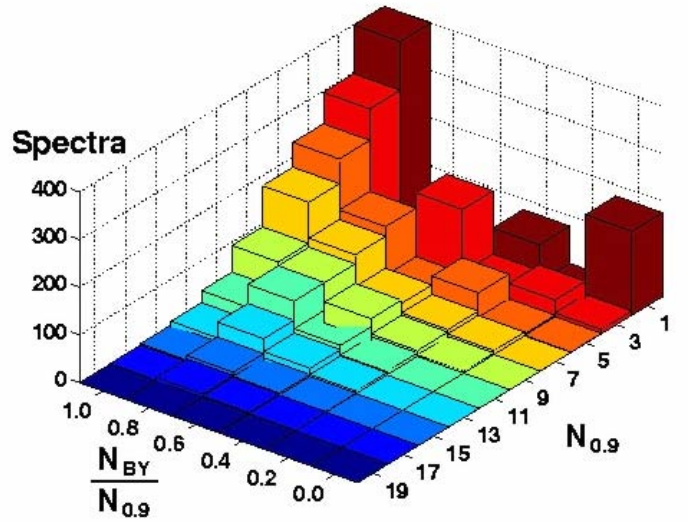


Fig. 2-1. Distribution of correctly predicted *b* and *y* peaks in the individual MS spectra demonstrates robustness of Probability Profile Method (PPM). $\mathbf{N_{0.9}}$ represents the number of the confidently predicted *b* and *y* peaks (P_{by} > 0.9), $\mathbf{N_{by}/N_{0.9}}$ represents the fraction of correctly predicted peaks.

We have made substantial progress this quarter in our molecular machines effort including a breakthrough in computational analysis tools for mass spectrometry data through the development of the Probability Profile Method demonstrating reliable ion identification in raw MS spectra using computational methods for the first time. This algorithm opens perspectives for a further development of whole range novel tools for MS analysis. In other work, we improved the Rosetta program to treat long loops, a capability important for several aspects of protein docking and developed and published a novel algorithm, KeyGeneMiner, for identification of genes responsible for the oxygenic photosynthetic process. This achievement lays an important foundation for selecting corresponding proteins and investigation of their interactions in this pathway. We also implemented a new optimization algorithm for our `LAMMPS` molecular dynamics code to enable multiprocessor parallel tempering to efficiently sample a wide range of temperatures to obtain equilibrated molecular structures. We carried out several simulations of Met-enkephalin, chosen as a test case because it is a representative of the short peptides to be used in our phage display experiments (Subproject 1) and because results can be compared to previously published tempering simulations of the same peptide (using different force fields). Finally, we enhanced our docking code, `PDOCK`, by integrating it with the `Coliny` optimization toolkit to access a wide range of global optimizers with a simpler user interface motivated by the need to model peptides with a mixed-domain representation for both continuous backbone torsion angles and discrete choices of side chains from a rotomer library.

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).

Oak Ridge National Laboratories, Sandia National Laboratories, Lawrence Berkeley National Laboratories, Los Alamos National Laboratories, University of California San Diego (Scripps), University of California Riverside, University of Michigan, University of Illinois Urbana/Champaign, National Center for Genome Resources, Molecular Sciences Institute, Joint Institute for Computational Sciences, University of Missouri.

Accomplishments

Computational Methods for Mass Spectrometry: Tandem Mass Spectrometry (MS/MS) is a workhorse of the modern high-throughput proteomics. At the present time tandem MS is employed for tasks ranging from characterization of protein complexes to whole cell proteome analysis. As discussed in previous reports, MS is one of a few feasible technologies for high-throughput characterization of molecular machines in the cell for several reasons, 1) MS is capable of providing useful information about multiunit assemblies, 2) it works with very small sample sizes, and 3) it picks up complexes in the solution state. In addition, chemical crosslinking techniques can be used to capture rare or transient protein complexes. Finally, many of the most difficult challenges for protein identification from MS spectra are algorithmic or computational.

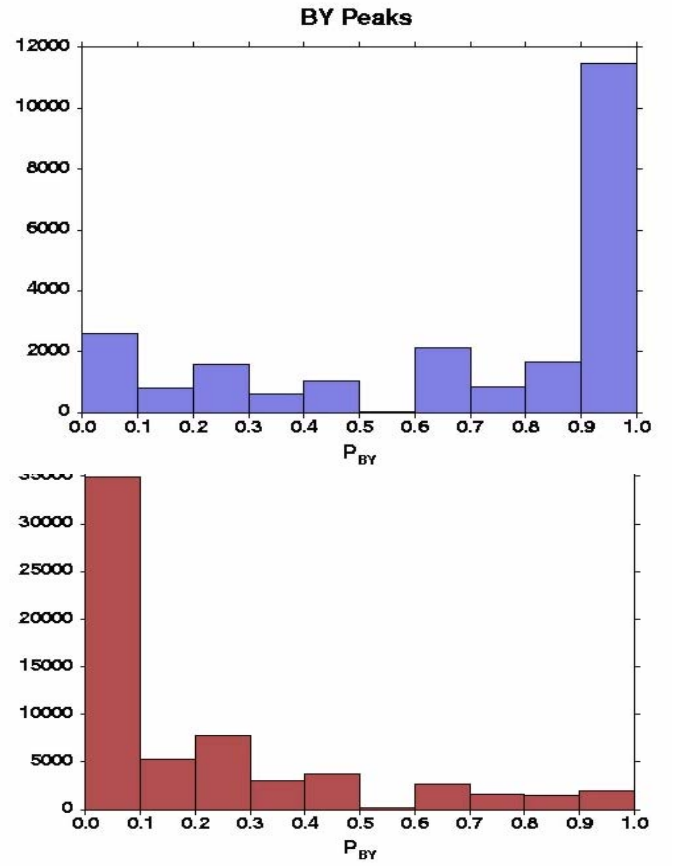


Fig. 2-2. Probabilities to belong to the category of b and y peaks are calculated separately for two set of peaks: a) actual b and y peaks (verified by SEQUEST analysis), b) actual r peaks - peaks of a comparable intensity which are not classified as b or y ions. Note the scale difference - there are 3 at least times more r peaks than b-y peaks.

In our June 2003 quarterly report we discussed the Probability Profile Method (PPM) for reliable identification of individual MS peaks. Our progress this quarter includes implementing and testing the PPM approach discussed as follows. We collected the frequencies of the satellite peaks observed for positively identified *b* class, *y* class, and *r* class (those otherwise not defined as *b* or *y*) peaks at relative positions +1, -17, -18, and –28 Da while including in our counts only those peaks (*b*, *y*, and *r*) which did not have another peak at –1 Da. This approach accomplishes two things, 1) it excludes from analysis approximately 23% of *r* peaks yet only 3% of *b* and 1% of *y* peaks, and 2) it yields much sharper observed differences at +1, -17, -18, and –28 Da. This is natural as isotope peaks have the same preferences for fragment formation as their primary ions, and inclusion in frequency measurements dilutes the discriminative power of the selected features. Next, as a comprehensive test of the method, we classified peaks into two categories: BY ("noble" *b* and *y* ions together) and R (rest of the peaks of comparable intensity) as presented in Figs. 2-1, 2-2, and 2-3. Two $\mathbf{P_{by}}$ distributions (defined as the computed probability of belonging to the BY category) are plotted in Fig. 2-2: for actual BY peaks (Fig. 2-2a) and for actual R peaks (Fig. 2-2b). Using this information we can answer an important question: how well does our probability function estimate the odds of finding BY peaks in the range of the obtained $\mathbf{P_{by}}$ values? We present the computed fraction of the real BY peaks for each 0.1 interval of the calculated $\mathbf{P_{by}}$ from 0 to 1 in **Fig. 2-3**. Note that in the most important top interval, $\mathbf{P_{by}}$ from 0.9 to 1, 85% of the predicted peaks are the actual BY peaks. In addition, we see monotonic correlation between the coverage of the individual spectra and, in the top intensity bin, we found 12,000 peaks within this interval (Fig 2-2). The number of high probability peaks discovered and

the quality of the predictions is presented in Fig. 2-1. The quality is estimated as a fraction ($\mathbf{N_{by}/N}$) of the correctly predicted peaks to all peaks with P > 0.9. For 1237 spectra (more than half of the total set) the quality is 1, which means that *all* of the predicted peaks are the *b* and *y* ions, the ideal situation for *de novo* sequencing.

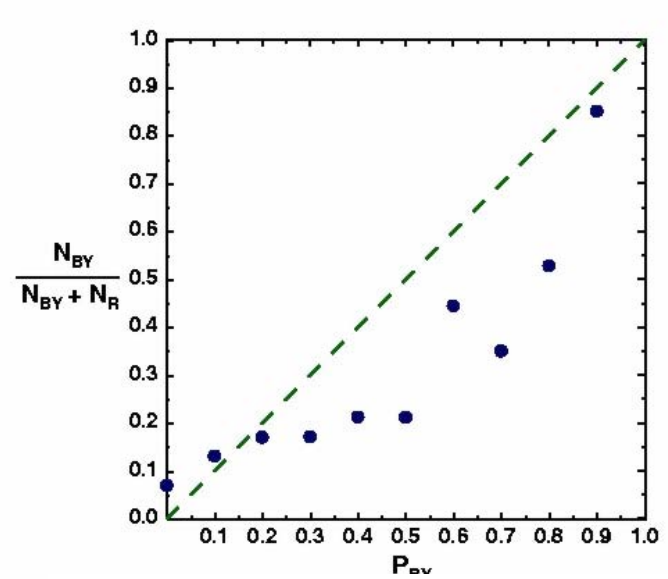


Fig. 2-3. Observed fraction of the actual BY peaks versus corresponding values of the calculated $\mathbf{P_{by}}$ probability. While data points are off the diagonal, the most important one (at 0.9 probability) is close.

Identifying Key Genes for Specific Biochemical Pathways: Identification of genes that are responsible for these processes is an essential step toward understanding their genetic and biochemical basis. However, such a task presents an enormous challenge for both experimental and computational scientists because of the complexity of gene networks involved in these processes. For experimental scientists, it is prohibitively expensive and labor-intensive to knock-out all possible genes to study particular cellular processes. In addition, function redundancies due to duplicated genes and alternative pathways may cause further complications in this analysis. Even for computational scientists, the evaluation of all possible candidate gene combinations is beyond current computational feasibility. Therefore, algorithms to identify critical components in these complex gene networks with high sensitivity and specificity are needed and represent a key goal of this project.

Current approaches to this problem are based on genome comparative analysis and ortholog-based genome comparative analysis but have been limited applicability due to a number of reasons discussed in our paper (Yu, Geist, Ostrouchov, Samatova, 2003). We have developed a novel algorithm (`KeyGeneMiner`) for the identification of "key" genome features responsible for a particular biochemical process of interest. The central idea behind our algorithm is that individual genome features (or their combinations) are identified as "key" if the discrimination accuracy between two classes of genomes with respect to a given biochemical process is sufficiently affected by the inclusion or exclusion of these features. The genome features are defined by high-resolution gene functions. The discrimination procedure utilizes the Support Vector Machine (SVM) classification technique. Changes in classification accuracy in response to addition or deletion of genome features measure the significance of these features.

We have applied this method to identify key genes responsible for the oxygenic photosynthetic process. Our algorithm identified 126 out of 8,254 highly confident key genome features. Twenty-seven of the features occur only among oxygenic photosynthetic genomes. The majority of these genome features (20 out of 27) are already documented as functions specific to the target process. In addition, we explored clustering of genes based on their co-occurrence on the genome. Such genome context analysis plays an essential role in protein function annotations. We discovered five clusters of genes that either correspond to the predicted key genome features or are known to be the protein components of the oxygenic photosynthetic process. They are highly conserved across multiple cyanobacterial organisms. Thus, we showed that `KeyGeneMiner` approach is able not only to identify the well-known components in the

oxygenic photosynthetic process, but also to discover novel candidate genome features that are completely unknown, even hypothetical proteins.

Developing of Loop modeling Protocols in Rosetta: Developing and extending the Rosetta method is an important part of our protein complex modeling strategy, which eventually will include Rosetta and Rosetta-like tools, as well as Monte-Carlo optimization in dihedral angle space (Internal Coordinate Mechanics) and molecular dynamics simulations. During this quarter we made advances in loop modeling, a problem that arises in two important contexts of protein binding. First, when two proteins bind, portions of their backbone structures may undergo significant rearrangements in the contact region. Thus a methodology for fixing the majority of a protein structure while remodeling short loops in an *ab initio* fashion is essential for testing possible rearrangements on the binding interface. Second, for some protein targets exact structures may not even be known, hence homology modeling combined with *ab initio* modeling of the structurally variable regions is required. This is especially important for our project, as we will have to rely on homology models for almost all *Synechococcus* proteins, at least in the short term.

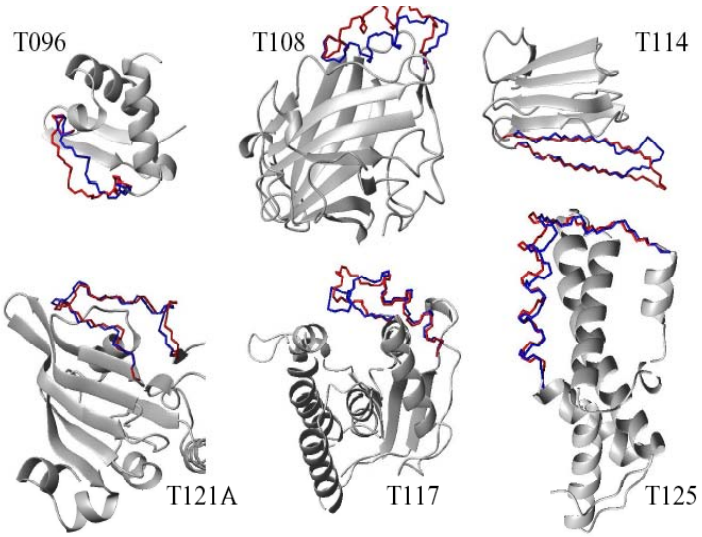


Fig. 2-4. Examples of loops reconstructed by Rosetta. The templates and loops were peer-selected instances that actually occurred in the blind structure prediction experiment known as CASP.

In collaboration with Carol Rohl (UC Santa Cruz) we have developed a scalable loop-modeling algorithm. Previously the best-of-breed loop-modeling strategies have been shown to perform poorly for long loops. These strategies are either based on exhaustive conformational search or selection of loops extracted whole from other proteins. Both of these approaches worsen as the loop length increases the complexity of the search space, and they begin to fail dramatically for inserts extending further than 6 to 12 residues. Our approach is to treat the loop modeling as just another *ab initio* modeling problem, and thus it opens an opportunity to scale it up to insertions of an entire domain size. The measured performance on test sets demonstrates that our method works about as well as the conventional methods on short loops and performance does not decline sharply as the loop length increases. (Rohl, Strauss, Baker, 2003)

The primary limitations are the following: loop modeling is a special problem for Monte Carlo searches because the structural constraints of the fixed endpoints of the loop are poorly compatible with random fragment insertion. To work within theses constraints we have developed a series of protocols for local move generation whose details have been submitted for publication (Rohl, Strauss, Misura, Baker, 2003). In Fig. 2-4, we show actual loop modeling problems taken from targets in the CASP experiments. The colored regions are the loops grafted on to a fixed backbone, (red is the correct structure, Blue is the Rosetta loop model). The results are promising and already of a useful accuracy.

Within a month we expect to have a public version of a Rosetta structure prediction server on-line. This will run on a 90 CPU Xeon cluster. It will provide fully automated structure predictions for both *ab initio*

model targets as well as for homology modeling using a web-based submission form. The homology modeling is a hybrid of *ab initio* modeling, a proprietary alignment protocol, and a meta-modeling approach to templates selection. In this approach, many other prediction servers are first queried for their best estimates of the template structure, then the sequence is re-aligned on these templates for a better structural match, finally, if there are variable structural regions in the models (inserted loops) these are modeled by *ab intio* methods. This system will be available both to the participants of this GTL project as well as to the public. Our collaborators, David Kim and Dylan Chivian (University of Washington) are largely responsible for this newest revision of the server.

Modeling MutL-MutS Molecular Machine: During this report period we started modeling simulation of the MutL-MutS bacterial molecular machine. The DNA mismatch repair (MMR) system is responsible for recognition and repair of errors during DNA replication in cells. In bacteria MMR malfunctions can confer hypermutator phenotypes, and in eukaryotes inactivation of the corresponding homologous genes leads to genome-wide instability. The MMR reaction pathway is a complicated process, involving multiple protein-protein interaction: MutS, MutL, endonuclease MutH, polymerase, lygase, and other components. While the MutL-MutS complex structure has long been desired due to the fact that its formation is the initiation step for other events, our interest in this problem is three-fold: 1) it represents a model molecular machine shown to be extremely important for proper functioning of bacteria cells, 2) it is a suitable problem to enable the comparison of our knowledge-based predictive tools (developed by Nagiza Samatova et. al., ORNL, Subproject 2) with modeling results, and 3) it presents an ideal opportunity to develop and optimize docking protocols while exploring novel ways to accelerate them.

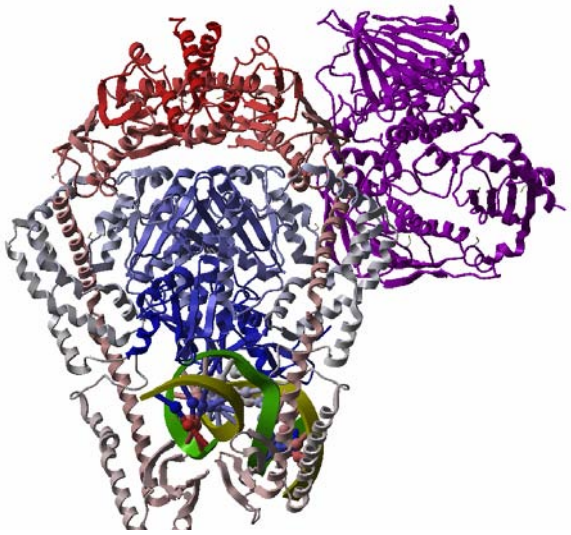


Fig 2-5. MutL-MutS complex DNA shown by green ribbon.

Direct experimental studies of the MMR machinery are extremely difficult, as they involve large protein-protein complexes that are transient in nature. At the same time isolated MutS and MutL homodimer structures have been solved by X-ray crystallography. For these reasons, high performance computational simulations could well be the only feasible way to elucidate structure of MutL-MutS complex and identify key residues important for the complex formation.

Extensive protein-protein docking simulations of MutL-MutS complex were conducted by Internal Coordinate mechanics (ICM) program with a pseudo-Brownian rigid body optimization step followed by a Biased Monte Carlo optimization of the side-chains interactions (flexible docking). Our immediate aim will be to develop general protocols of protein docking, which could be later applied for modeling of other bacterial machines, and test novel ideas to accelerate the process. This approach will enable a powerful coupling between the Rosetta approach to modeling of individual protein domains and MD simulation in explicit solvent.

Molecular Dynamics Simulation of RuBisCO Protein and Peptide:. Integration of SHAKE (bond and angle constraints) and PPPM (particle-particle particle-mesh for long-range Coulombics) solvers into the newest version of our molecular dynamics code LAMMPS was completed this quarter. In parallel tempering mode (replica exchange), the new LAMMPS can simulate multiple replicas of a solvated peptide (or protein) at different temperatures on different subsets of processors. Monte Carlo temperature swaps are performed periodically between neighboring ensembles to enable more efficient sampling of the conformational space accessible to the peptide. Multi-nanosecond runs of a peptide 5-mer Met-enkephalin have been run on 64 processors of machines at ORNL and SNL using the CHARMM force field and PPPM solver.

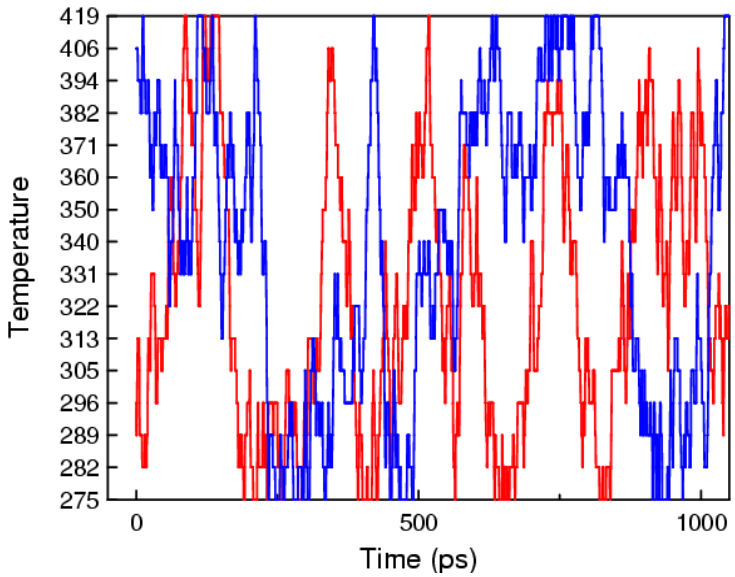


Fig. 2-6. Temperature profiles of 2 replicas (red and blue) for the first nanosecond of a parallel tempering simulation. The full simulation used 16 replicas ranging in temperature from 275K to 419K.

Temperature profiles for 2 of the 16 ensembles in the simulation are shown in Fig. 2-6, indicating how the tempering algorithm enables each ensemble to sample a wide range of temperatures. Met-enkephalin was chosen as a test case because it is a representative of the short peptides to be used in our phage display experiments (Subproject 1) and because results can be compared to previously published tempering simulations of the same peptide (using different force fields). Post-processing of the simulation data is currently underway as are simulations of longer peptides (15-20 mers) which undergo more interesting structural changes.

Molecular dynamics simulations of a dimer of large subunits from *Synechococcus* RuBisCO were setup and run using the CHARMM 27 force field in LAMMPS. X-ray structure 1RBL from the protein data bank was used as the starting coordinates for the large subunits. We modified the force field to include carbamylated lysine as well as the eneolized ribulose 1,5-bisphosphate and CO2 substrates.

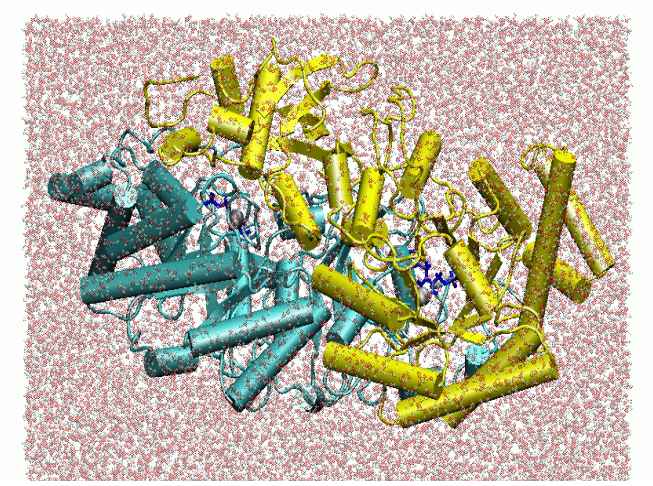


Fig. 2-7. Snapshot of a molecular dynamics model of RuBisCO, containing approximately 52,000 atoms. The protein dimer contains 2 large sub-units of *Synechococcus* RuBisCO (blue and yellow); the catalytic binding sites for CO_2 are shown as small grey spheres; water molecules are shown in purple.

Simulation setup included choosing protonation states for the residues and including an explicit ionic solvent for charge neutralization of the simulation cell. The final simulation cell (75x75x113 Angstroms) consists of ~52,000 atoms, with 2 catalytic sites - see Fig. 2-6. Thus far the system has been simulated for about 0.7 nanoseconds (350,000 timesteps) on 64 processors of an SNL parallel machine. Post-processing of the RuBisCO simulation data is proceeding. We have examined the protein secondary structure over the course of the simulation and have observed good stability of the known structure for the entire simulation time. We have also examined the backbone dihedral angles of the ribulose 1,5-bisphosphate molecules, which show rigidity due to the tight coupling with the enzyme. Further post-processing analysis will focus on interactions of key RuBisCO residues with the substrate molecules.

Metal ions are essential to the specific structure and function of RuBisCO. For example, the active site of RuBisCO contains Mg2+ and can accommodate many other divalent ions such as calcium, nickel, manganese, iron and copper. Different metal ions at the active site change the function of the enzyme. Prior to studying cation binding to the RuBisCO active site, we have modeled their properties in bulk water using a multiscale quantum/classical approach validated by *ab initio* molecular dynamics simulations and experimental data. A quantum description in the local region of the ion is required because of the dominance of local chemical effects (changes in electron distributions).

Metal ions are essential to the specific structure and function of RuBisCO. For example, the active site of RuBisCO contains Mg2+ and can accommodate many other divalent ions such as calcium, nickel, manganese, iron and copper. Different metal ions at the active site change the function of the enzyme. Prior to studying cation binding to the RuBisCO active site, we have modeled their properties in bulk water using a multiscale quantum/classical approach validated by *ab initio* molecular dynamics simulations and experimental data. A quantum description in the local region of the ion is required because of the dominance of local chemical effects (changes in electron distributions).

Peptide-protein docking will be performed with `PDOCK`, our docking code that uses structure-based algorithms to dock peptides to proteins of known structure and evaluate their potential binding affinities. We have enhanced `PDOCK` by integrating it with the `Coliny` optimization toolkit. `Coliny` is a revamped version of the `SGOpt` library, which we have also integrated into `PDOCK`. `Coliny` offers a wider range of global optimizers with a simpler user interface. The use of `Coliny` is also motivated by ongoing efforts to model peptides with a mixed-domain representation: continuous backbone torsion angles and discrete choice of side chains from a rotomer library. Our next task is to model peptides with this representation and to assess the scalability of docking with this formulation. We are working with Tony Martino (Subproject 1) to identify candidate peptide sequences that are involved in protein-protein interactions relevant to *Synechococcus*.

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 (8/03). During reporting period we made a significant advance with Probability Profile Method – novel method for peak identification in MS spectra. In the last quarterly report we reported implementation of the relevant residual dipolar coupling methods as well.

Aim 2. Develop parallel tempering technology, all-atom docking models for flexible peptide chains, CB-MC techniques (04/03). We successfully implemented the first version of parallel tempering in the `LAMMPS` code and all-atom docking methods for peptides.

Aims 3, 4. Develop categorical analysis tool combing several genome context data sources for analysis of protein-protein interaction (08/03). Implementation of `PICCUP` method was reported in the last quarterly report, and we have had a journal paper accepted for publication.

Aims 3, 4. Create catalog of proteins in *Synechococcus* that are relevant to specific metabolic pathways, (including SMR and ABC transporters, channels) (08/03). The first implementation of a new method, `KeyGeneFinder`, has been completed.

Collaboration With Others

In development of mass-spec analysis capabilities we actively collaborated with ORNL-PNNL GTL project, "Center for Molecular and Cellular Systems": Edward Uberbacher, Gregory Hurst, Frank Larimer, Tema Fridman and Jane Razumovskaya participated in work or contributed to the discussions. We also worked extensively with Tony Martino, SNL/CA (Subproject 1 PI) and Hashim Al-Hashimi, U Michigan (Subproject 1) on the exploration of molecular machinery with NMR.

We started collaborating with Brian Palenik, UCSD/Scripps Research Institute (Subproject 1) on applying `KeyGeneMiner` to identify key genes specific to marine genomes. We are in the process of generating the data required for comparative genome analysis. Computational molecular biophysics collaborations include docking work with CU Denver, U Padua, MIT, U Nottingham, on MD simulation work with Tom Woolf at Johns Hopkins.

Collaborators on Rosetta improvement include members of the David Baker lab, U Washington: David Kim and Dylan Chivian, on the `Robetta` server port. Modeling of MutL-MutS was done with Ruben Abagyan laboratory at UCSD/Scripps Research Institute.

We also collaborated with Carol Rohl, UC Santa Cruz to develop a scalable loop-modeling algorithm.

Publications and Presentations

1. Fridman, Tema, Day, Robert, Razumovskaya, Jane, Xu, Dong, Gorin, Andrey, "Probability Profiles - Novel Approach in Mass Spectrometry De Novo Sequencing". Poster at the *IEEE Computer Society Bioinformatics Conference,* Stanford, CA, August 2003.

2. Rohl, Carol A., Strauss, Charlie E. M., Chivian, Dylan, Baker, David "Modeling Structurally Variable Regions In Homologous Proteins With Rosetta" *Proteins,* Submitted, August 2003.

3. Rohl, Carol A., Strauss, Charlie E. M., Misura, Kira M. S., Baker, David "Protein Structure Prediction Using Rosetta", *Methods Enzymo*l, Submitted, August 2003.

4. Strauss Charlie E.M., "Sequence to Structure using Shortle's Propensities." *Rosetta Developers Conference*, Leavenworth, Washington, Invited, Aug 8,9 2003.

5. Yu, Gong-Xin, Geist, Al, Ostrouchov, George, Samatova, Nagiza, "An SVM-based Algorithm for Identification of Photosynthesis-specific Genome Features", *Proceedings of the IEEE Bioinformatics Conference*, Stanford, CA, August 11-14, 2003.

6. Yu, Gong-Xin, Geist, Al, Ostrouchov, George, Samatova, Nagiza, "An SVM-based Algorithm for Identification of Photosynthesis-specific Genome Features", *Presentation at the IEEE Bioinformatics Conference*, Stanford, CA, August 11-14, 2003.

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

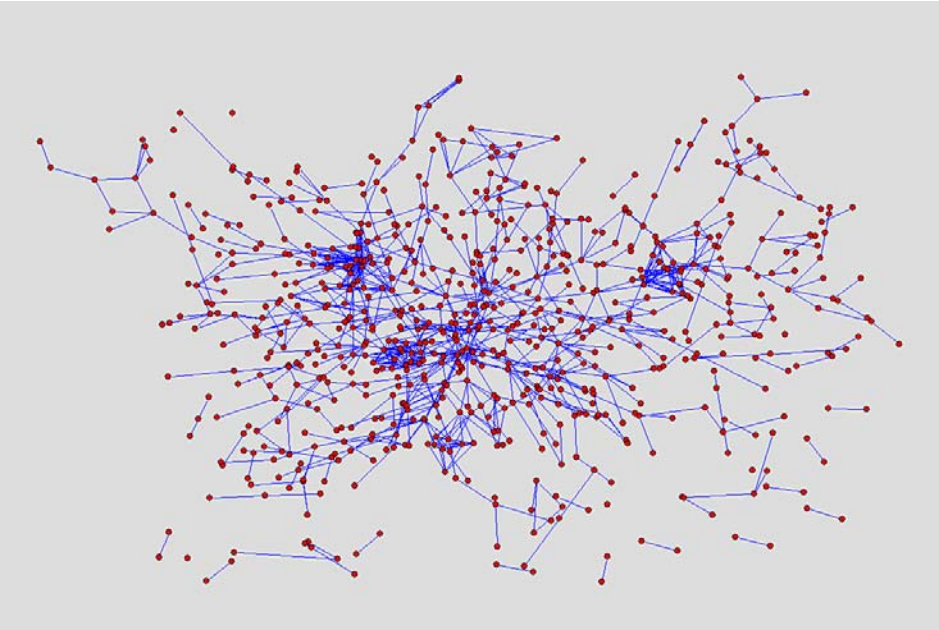


Fig. 3-1. Protein-protein interaction network with 722 proteins involved in 963 interactions in *Synechococcus* WH8102, derived data mining and information fusion.

Progress continues in our effort to develop new and effective protocols for tackling the challenge of systematic characterization of regulatory pathways. In particular, we have developed and prototyped several new computational methods including a preliminary version of a computational pipeline for interpretation of multiple types of biological data. Ultimately, this effort will form part of a computational framework for biological pathway inference including the prediction of operon structures and the identification of regulatory binding sites in *Synechococcus.* We have also applied our methods to develop several predictions for *Synechococcus* WH8102 including a signaling/regulatory network for the phosphorus assimilation pathway and a genome-scale protein-protein interaction map. Our efforts to develop new microarray analysis capabilities and apply them to *Synechococcus* continue; we have completed the print design of the whole genome *Synechococcus* microarray to be fabricated by TIGR and have analyzed data from 38 DNA-DNA and RNA expression microarrays for the 250 gene *Synechococcus* microarray from TIGR with our methods. Initial results indicate that TIGR's microarray process needs further investigation to achieve the desired reproducibility.

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).

Oak Ridge National Laboratories, Sandia National Laboratories, Lawrence Berkeley National Laboratories, Los Alamos National Laboratories, University of California San Diego (Scripps), University of California Riverside, University of Michigan, University of Illinois Urbana/Champaign, National Center for Genome Resources, Molecular Sciences Institute, Joint Institute for Computational Sciences, University of Missouri.

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 function predictions, protein-protein interaction predictions, operon structure prediction, transcription factor binding site identification, and statistical analysis of microarray gene expression data. While these computational capabilities can be used for more general purposes, we have used these tools to extract information for our pathway predictions.

In particular, we have developed and prototyped several new computational methods including 1) a preliminary version of a computational pipeline for interpretation of multiple types of biological data, including genomic sequence data, two hybrid data, microarray data, etc., as part of the computational framework for biological pathway inference, 2) a computer program for prediction of operon structures for cyanobacterial genomes, and 3) a new method for identification of regulatory binding sites in *Synechococcus.* We also applied these methods to *Synechococcus* WH8102 to make several predictions for experimental validation including a signaling/regulatory network for the phosphorus assimilation pathway and a genome-scale protein-protein interaction map. Work is also underway to develop predictions for two additional signaling/regulatory networks for nitrate and carbon assimilation in *Synechococcus* WH8102.

We continue to make good progress in our efforts to develop and refine microarray analysis capabilities and apply them to *Synechococcus*. To this end, we have completed the print design of the whole genome *Synechococcus* microarray to be fabricated by TIGR. Our design includes genes repeated six times in a statistically designed fashion and includes large numbers of control spots in each block of the microarray. We also developed and tested computational implementations of maximum likelihood classification methods for incorporating error covariance estimates of real microarray data to simulated gene expression data as well as feature extraction methods for comparing their gene selection ability using simulated microarray data. We also analyzed data from 38 DNA-DNA and RNA expression microarrays for the 250 gene *Synechococcus* microarray from TIGR with our methods. Initial results indicate that TIGR's microarray process needs further investigation to achieve the desired reproducibility.

Progress Towards Milestones

Aim 1. Complete series of designed experiments to characterize error structure associated with measuring replicate arrays; generate, code, and test computational methods incorporating error covariance estimates of real microarray data (10/03). Based on our analyses of data from replicate cDNA microarrays, we have proposed a print design for the full *Synechococcus* genome microarray to TIGR. TIGR is under contract to Sandia to develop and print these microarrays. The design concepts include using a common control spot pattern across blocks as well as gene replicates spread across blocks of the array. The print design provides 6 replicates of the 2496-member gene set as well as replicates of 88 control spots in each of the 48 blocks. TIGR will select the specific control spots to be used in the whole genome microarray.

Aim 2. Generate simulated microarray data with realistic error structure and use simulated data to test sensitivity of various clustering and classification algorithms; implement our new clustering algorithms for gene expression data (7/03). Work has continued relating to the development of a realistic simulation tool used to compare algorithms for identifying highly expressed genes. The tool now allows one to include correlated errors that might be expected when performing the hybridization of microarrays in batches. A variety of univariate and multivariate gene selection algorithms were applied to at least a dozen simulated data sets of various relative gene signal expression levels with heteroscedastic

errors with and without correlated errors present. The methods tested included t-tests, Bootstrapped Analysis of Variance (ANOVA), Wilcoxon method with boosting, shuffled Fisher forward feature selection, and support vector machines with recursive feature extraction (SVM-RFE). In addition, software has been written to implement a jackknife multivariate feature selection method, but it has not yet been applied to the simulated data. The results of the application of these various gene selection tools to the realistically simulated data demonstrates that the Bootstrapped Analysis of Variance (ANOVA) and Wilcoxon method with boosting both yielded comparable results which far exceeded the success of the other gene selection tools. In addition, the presence of correlated noise was harmful to the gene selection when the gene expression levels were low.

The use of maximum likelihood principal component regression (MLPCR) methods for supervised clustering of the data have been programmed in Matlab and tested with our realistically simulated microarray data. Comparing the MLPCR method with standard principal component regression (PCR) applied to the simulated data with heteroscedastic errors confirms that MLPCR has a significant improvement in supervised clustering for a two class problem.

Aim 3. Develop binding-site identification methods and implement the methods in a computer program (9/03). We have developed a novel computer program, based on our minimum spanning tree-based data clustering framework, for identification of regulatory binding sites. This program is currently being used for identification regulatory binding sites relevant to the process of phosphate assimilation, carbon fixation, and other important biological processes in *Synechococcus*.

Aim 5. Implement basic toolkit for database search (10/03). We have predicted a signaling/regulatory network model for the phosphorus assimilation process in *Synechococcus* WH8102, and are in the process of building two additional network models for carbon and nitrate assimilation processes. We have gained great experience for piecing together pieces of information derived from different sources and putting them to infer a pathway model. While we will continue to work on individual pathway model predictions, we start to build computational inference capabilities for automated pathway inference and apply it to many signaling pathways.

Aim 6. Refine approaches for scanning and analyzing our DNA microarrays; provide slides that we have scanned for inter-lab 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, develop programs to read and begin to understand the relevant text (08/03). We have been in close contact with our experimental collaborator Brian Palenik, UCSD/Scripps Research Institute (Subproject 1) resulting in significant benefits to our pathway inference processes research.

Collaboration With Others

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

David Haaland, SNL (Subproject 3) had numerous discussions with Ian Paulsen at TIGR in Rockville, MD to discuss their *Synechococcus* microarray data and experimental procedures. He also discussed with Ian Paulson the proposed whole genome *Synechococcus* microarrays that TIGR will generate and hybridize for our project. Both David Haaland, SNL (Subproject 3) and Ed Thomas, SNL (Subproject 3) visited TIGR in early May to discuss microarray results and to make suggestions for improved microarray data.

Publications and Presentations

1. Su, Z., Dam, A., Chen, X., Olman, V., Jiang, T., Palenik, B., and Xu, Y., "Computational Inference of Regulatory Pathways in Microbes: an Application to the Construction of Phosphorus Assimilation Pathways in *Synechococcus* WH8102", *Proceedings of the 14th International Conference on Genome Informatics*, in press, Dec. 2003.

2. Su, Z., Dam, A., Chen, X., Jiang, T., Palenik, B., and Xu, Y., "Computational Inference of Regulatory Pathways in Microbes in *Synechococcus sp"*, *Genome Research*, Submitted, 2003.

3. Xu, Ying, "A New Algorithm for Computational Inference of Microbial Regulatory Pathways", Biology Department, Georgia Tech, Nov. 2003.

4. Xu, Ying, "Inference of Regulatory/Signaling Pathways in a System Manner", distinguished lecture, *Computer Science Department, Alberta University, Canada*, Invited, Jan. 26, 2004.

5. Xu, Ying, "A Computational Framework for Inference of Regulatory and Signaling Pathways in Microbes", *14th International Conference on Genome Informatics*, Tokyo, Japan, Dec. 15th, 2003.

6. Xu, Ying, "Computational Inference of Signaling and Regulatory Pathways in Microbes", *Fudan Bioinformatics Workshop*, Invited, Dec. 18, 2003.

7. Xu, Ying, "Computational Inference of Regulatory/Signaling Pathways in Microbes", *Pathology Department, UTK*, Invited, June 5, 2003.

8. Xu, Ying, "Computational Inference of Regulatory/Signaling Pathways in *Synechococcus* WH8102", *ORNL/CRAY Biology Workshop*, Invited, May 9, 2003.

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

Systems Biology for *Synechococcus* Sp. (Subproject 4) Highlights

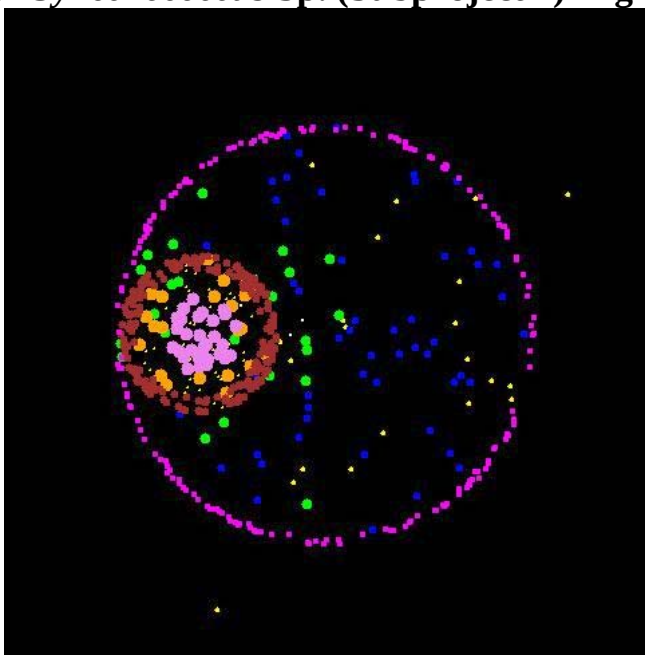


Fig. 4-1. A 2d snapshot from a 3d simulation of glucose production in a carboxysome of an idealized *Synechococcus* cell. The purple and red particles outline the cell and carboxysome boundaries; pink particles are carbonic anhydrase, orange are RuBisCO, green are glucose, blue are ribulose, small yellow particle in the cell are bicarbonate ions, small yellow particles in the carboxysome are CO2 molecules. The movie may be viewed at http://www.genomes2life.org/highlights/

We have completed the simulation framework for the discrete particle model and have begun to run model simulations that incorporate many of the relevant components of the carbon cycle. By viewing these simulations as movies we not only get a quantitative sense of how reaction rates and concentrations affect the simulation, but also get a more fundamental physical understanding of the underlying processes and how they are affected by the geometry of the cell. Among the bioinformatics highlights, we have discovered a rigorous mathematical method for making predictions regarding protein-protein interactions across related organisms using sequence data. We have also initiated work to develop computational tools to enable us to model the flow of carbon into the ocean and through the water towards the individual *Synechococcus* cells as the next step at the extra-cellular resolution level towards our hierarchical global carbon fixation modeling strategy and completed our first rudimentary implementation of our system model. This platform implements a discrete event model as the architecture for simulations encompassing disparate temporal scales. The architecture is designed to eventually allow sub-simulations that may take arbitrary wall clock time to complete, to exchange synchronized events with sub-simulations that may take negligible wall clock time. The first run prototype uses a test example to demonstrate our working solution to the disparate temporal scale problem.

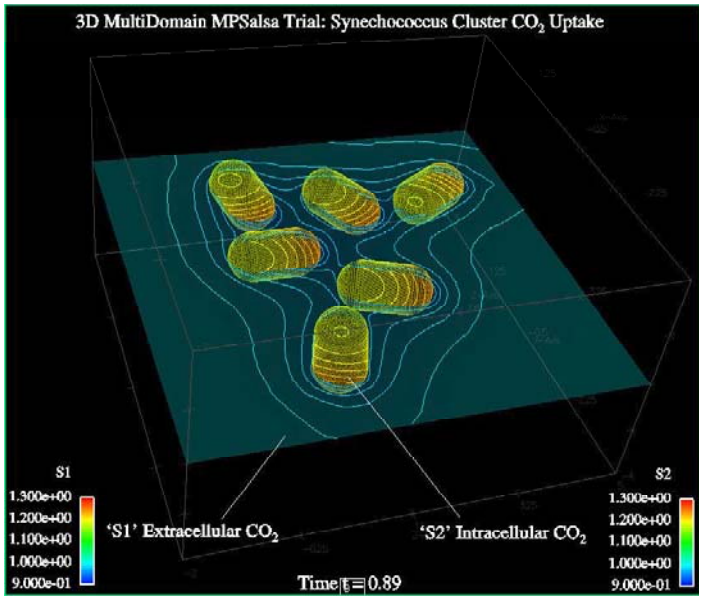
This work was funded in part or in full by the US Department of Energy's Genomes to Life program (www.doegenomestolife.org) under project, "Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org).

Oak Ridge National Laboratories, Sandia National Laboratories, Lawrence Berkeley National Laboratories, Los Alamos National Laboratories, University of California San Diego (Scripps), University of California Riverside, University of Michigan, University of Illinois Urbana/Champaign, National Center for Genome Resources, Molecular Sciences Institute, Joint Institute for Computational Sciences, University of Missouri.

Accomplishments

Protein Interaction Network Inference and Analysis: Related to Aim 1, the main accomplishments of the quarter have been 1) to establish that the scale-free characteristic of biological networks is not a topological property but rather a property of the process that creates the network, and 2) to develop a novel bioinformatics technique to predict protein-protein interactions across related organisms.

Discrete Component Model: We have continued development of a particle-based model of protein/protein interactions within cells. Individual particles represent protein molecules or protein complexes. These particles diffuse through the simulation geometry via Brownian motion pairing with nearby particles using Monte Carlo rules to react according to an input list of possible chemical reaction equations. This quarter we added cellular compartments to the simulation to represent membranes and intra-cellular compartments, such as the nucleus or other vesicles within the cell. Particles of various species are assigned to compartments and permeability rules govern how particles move from one compartment to another. In Fig. 4-1 we show how these features can be used to generate a simple model of a portion of the carbon fixation process in *Synechococcus* (the movie can be viewed at http://www.genomes-to-life.org/highlights) In this case, the cell contains a single carboxysome, and its carbonic anhydrase proteins convert bicarbonate ions (which diffuse into the carboxysome from the cell cytoplasm) into CO_2, which then gets converted by activated RuBisCO (RuBisCO plus ribulose) into glucose. We intend to use coarse-grain models such as this to deduce the sensitivity of glucose production (the product of carbon fixation) to parameters that control the reaction and diffusion properties of the various protein reactants.



Continuum System Models for Synechococcus: This quarter we embarked upon effort to modify the `MPSalsa` code to handle multi-domain problems, allowing simulation of colonies or plaques of *Synechococcus* and their impact on surrounding environmental concentrations of carbon dioxide. This will be fundamentally important to the question of how the carbon fixation phenomena at lower length and time-scales are incorporated into the hierarchical model of the global carbon fixation process. We also continue to test operator-splitting solution methods for biologically relevant source terms in our reacting-flows solver `MPSalsa`, with the aim of increasing solution efficiency while maintaining acceptable approximation accuracy.

Hierarchical Simulation Platform: Our major accomplishment this quarter was implementing the disparate temporal scale aspects of the high-level Design Document into C++ code using standardize design patterns. This required architectural, software engineering, and coding expertise. Designing and implementing a solution to the "disparate temporal scale" problem was an important accomplishment because it is both central to the platform itself as well as of general interest to advances in simulation science. Our solution uses a MasterClock broker on top of an *Observer* pattern to coordinate discrete time-stamped events from sender to receiver, such that events are generated as driven by CPU-time

constraints, yet are delivered in monotonically increasing simulation time. Under this model, receivers react to events as they are received and are guaranteed that no event ever occurs "in its past" regardless of any CPU/simulation time scale disparity between sender and receiver. By coordinating *TimeAdvanceRequest* events, the simulation always advances to next greatest possible time, thus avoiding any necessity to time-slice on the smallest time unit. Thus the platform can coordinate sub-simulations that take both milliseconds and years without needing to time slice a year into $10^3 \times 60 \times 60 \times 24 \times 365.25 = 3.15576 \times 10^{10}$ milliseconds.

Progress Towards Milestones

Aim 1. Develop graph theoretical tools for network analysis. Use tools to characterize the scale-free nature of protein interaction networks and publish analyses results on existing protein interaction networks (9/03). We have performed a mathematical analysis of scale-free graphs by considering the random removal of edges from graphs with power-law degree distributions. We expected that this process would yield, on average, smaller scale-free graphs. We discovered, however, that random removal of edges from a power-law graph is not expected to preserve the power-law, and that the deviation is small enough that it is undetectable using computational simulations. Our result on power-law graphs suggests that it is not the network itself that is scale-free, but rather the process that forms the network. In other words a scale-free process will form networks of any size, all satisfying the power-law, but an individual network is not by itself invariant to changes at different scales. A paper summarizing these findings has been submitted to Internet Mathematics.

We have also made progress during this quarter toward proposing a protein network for *Synechococcus* WH8102. Our protein-protein in-silico prediction technique presented in the last quarter, initially devised for phage-display data, has been generalized and can now take as input 2-hybrid experimental data. Our method is based on the use of Support Vector Machines (SVMs) applied to amino acid sequence signature descriptors. Specifically we use a height-one product signature in order to describe protein pairs. This amounts to every possible pair of amino acid tri-mers for each pair of proteins. As there is a large number of such tri-mers and many pairs of proteins we were required to construct two novel functions for use with classifiers. In particular we developed a signature kernel, a signature product kernel, and a relation between the two that enabled us to actually implement the SVMs. The novelties of this method lie in the use of signature, the idea of the signature product, and the implementation of the method using the SVM methodology. We have tested our method using 2-hybrid data obtained with *Yeast* and *H. Pylori*. Our signature kernel method gave an 80% prediction accuracy on both datasets and our predictions on the entire proteome in both cases yielded networks which were scale-free and anti-correlated. The support vectors trained with *H. Pylori* data were then used to predict the possible pairwise protein-protein interactions for the entire to *Synechococcus* WH8102 proteome. Note that the genomes of *Synechococcus* WH8102 and *H. Pylori* have been completed and that they both belong to the Eubacteria family. While specific predicted interactions are being verified experimentally within Subproject 1, the general topology of the predicted *Synechococcus* network was found to follow a power law and an anti-correlation law.

Aim 2. Begin to test the code on yeast data and *Synechococcus* data (if available) (9/03). *Synechococcus* data from the literature has been studied to determine the particle numbers and typical reaction times when available. These parameters have been incorporated into the simulations and we are attempting to infer values for other reaction parameter that cannot be easily measured via experiment. Our project's discrete particle simulation team, comprised of team members from Sandia and The Molecular Sciences Institute, is beginning more extensive collaborations to prototype our simulation on a system that has much more complete data in order to test its ability to capture real systems accurately.

Aim 3. Start to perform reaction/diffusion simulations using preliminary boundary information to test the membrane/ion channel work against experimental data (3/04). Our work has focused on two elements of this objective, 1) code testing and implementation of novel, operator-splitting methods, and 2) modification of multi-domain problems representation for `MPSalsa` solution.

Operator Splitting Methods: Initial studies of smooth, continuous source terms demonstrate splitting methods can achieve equivalent accuracy of fully implicit methods. We have established the second-order accuracy for several of the splitting methods, however we are still working out the details for situations where the source terms include discontinuities. Comparisons between fully implicit second order methods and second order splitting methods show both methods provide only first order convergence of L2 norms against exact solutions (one-dimensional reaction-diffusion partial differential equations). We continue testing and literature searching for comparisons with other similar studies as regards the discontinuous source term. Therefore, we intend usage of continuous, smooth source terms in the initial application of operator splitting methods for relevant reactions in *Synechococcus*. This is not unreasonable given that most physiological reactions in nature are not discontinuous; however, we intend further study of this first-order convergence phenomenon due to the wide usage of such discontinuous reactions in computational biology, and the probability that experimental research may demonstrate threshold activations/inactivations of various relevant reactions in *Synechococcus*. Optimized implementation of splitting methods (improved speed, memory usage, etc.) awaits completion of such testing, and should prove straightforward when required.

Multi-Domain Solution in `MPSalsa`: `MPSalsa` was originally designed primarily for solution of reacting fluid flows with diffusive/thermal gradient transport within single domain; i.e., `MPSalsa` was not originally intended for solution of problems with interior boundaries and jump conditions for species' concentrations. In other words, a single node cannot have two values for the same species. This limits the type of problems accessible to `MPSalsa` to single domains where no jump-conditions of species' solutions permissible within the node list. Hence, we have studied alternative methods of handling this interior domain jump condition. The most straightforward approach utilizes an interim layer of elements between two interior domains to handle the jump-condition, and further provides a spatial region to localize transport reactions. For example, the transport mechanism for *Synechococcus* on its surface is presumably a protein structure embedded (and probably diffusing) in the membrane of the cell. This membrane and transport system is more amenable to simulation with a geometric and meshed representation of the membrane in the simulation domain. In order to accurately represent this transport, we convert between the more typical Neumann flux representation of such transport to a spatially localized reaction in the interim boundary region, or membrane. Initial one-dimensional tests show such a spatially localized reaction term can achieve acceptable accuracy if the interim region is relatively small compared to the overall simulation domain – similar to the relative size between the cellular membrane and the cell. This relative size limitation is evident from artifices introduced into the solution profile if the transport domain is too large. Further three-dimensional testing is underway, and we also intend to utilize the above referenced operator-splitting methods for these surface reaction representations of surface transport, since they are designed for such reaction/transport solution.

Aim 4. Finish first code implementation (9/03). Based on the Use Case, Requirements, and Design documents delivered in 06/03, we have implemented the first run prototype in C++. The deliverable consists of source code with comments that has been successfully compiled by GNU gcc and capable of executing a behaviorally demonstrable test case in a Solaris environment.

Collaboration With Others

We have provided a list of predicted the protein-protein interactions for *Synechococcus* WH8102 to Subproject 1 team members for experimental investigation/feedback. Specific interactions related to TPR domains are being checked experimentally using 2-hybrid experiments.

We have engaged and hired the expertise of Dr. Andy Belgrano at the National Center for Genome Resources (NCGR). Dr. Belgrano has worked on the NSF bio-complexity project and brings both oceanographic and marine bacterial biology expertise to the project.

We are starting the process of bringing all of the aims in Subproject 4 together as a whole by incorporating information from the discrete simulations into the continuum simulations.

Publications and Presentations

1. Churchwell, C. J., Rintoul, M. D, Martin, S., Visco D., Kotu, A., Larson, R. S., Sillerud, L. O., Brown, D. C., Faulon, J.L., "The Signature Molecular Descriptor. 3. Inverse Quantitative Structure-Activity Relationship of ICAM-1 Inhibitory Peptides," *J. Molec. Graph. and Model*, Submitted.

2. Faulon, J.L., Martin, S., "Using Graph Extended-Degree Sequence to Characterize Molecular Graphs and Biological Network," *American Math Society, Symposium on Computational and Mathematical Biology*, Boulder, CO, Oct. 2003.

3. Faulon, J.L., Martin, S., Visco, D. Kotu, A., "Probing Information Content in QSAR Analyses Using the Signature Molecular Descriptor," *American Chemical Society National Meeting*, New York, NY, Sept. 2003.

4. Gessler, D., Blanchard, J., "Use Case Document: Hierarchical Simulation Platform,", *National Center for Genome Resources*, June 2003.

5. Martin, S., Carr R.D., Faulon, J.L., "Random Removal of Edges from Scale Free Graphs," *Internet Mathematics*, Submitted.

6. Steadman, P., Gessler, D., "Design Document: Hierarchical Simulation Platform," National Center for Genome Resources, June 2003b

7. Steadman, P., Gessler, D., "Requirements Document: Hierarchical Simulation Platform," *National Center for Genome Resources*, June 2003.

Computational Biology Work Environments and Infrastructure (Subproject 5) Highlights

Fig. 5-1. We have established a data and computational infrastructure in use by the project include the hardware in this figure, the ORNL TORC computing cluster, an ORNL-based SUN oracle server as well as a recently acquired three terabyte storage server.

Our computational work environments and infrastructure effort made significant progress this quarter toward establishing a data and computational infrastructure for use by the project. To this end, we installed hardware (at ORNL), initiated development of tools for browsing protein interaction networks and identifying functional association links between proteins. We also designed a language for our Matlab-like bioinformatics work environment ("BiLab") which incorporates the scripting style of Matlab with a range of modern language features for larger scale and library style generic and object-oriented software development.

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 in use by the project. To this end, we installed (at ORNL) two Sun iForce clusters that share a common D1000 disk farm. One cluster is a high-performance Web Server, and the other is an Oracle database server. We also ordered a three TeraByte storage server on which to create a Model Organism Database (MOD) with all known data on *Synechococcus* and its close relatives. Other work included initiating the development of tools for browsing protein interaction networks, and algorithms for identifying functional association links between proteins via gene fusion, phylogenetic profiling, gene-order, and so on. Finally, our Subproject 5 team continues to maintain the project's web site, keeping calendars, publications, and job opportunities updated. We also updated the project's Electronic Notebook software and carried out (LBNL) all of the preparations to host the project's semi-annual meeting October 21-22, 2003, in Berkeley, CA.

Progress Towards Milestones

Aim 1. Integrate new methods and tools into an easy to use work environment. Creation of prototype electronic lab notebook that handles biological data types (9/03). We continue to develop a system for browsing protein-protein interaction networks. Interactions include both physical as well as functional links derived from various information sources or predicted using computational tools. Our tool will be web-based and will support various graph-based queries that are being developed at LBNL by Subproject 5 team members Frank Olken and Arie Shoshani. Each node and edge of a network will be annotated by various genomics and proteomics information such as the source of interaction (e.g., experimental, predicted), the type of tool, confidence levels, etc.

Aim 2. Complete design of general-purpose graph-based data management system (9/03). We added *k*-core queries, and graph characterization queries to our list of query types. We began work on context representation for bioprocesses (reactions). Contextual annotations are used to specify the circumstances under which reactions occur.

At the suggestion of Arie Shoshani, LBNL (Subproject 5) we have begun planning to add a visual, graph-based query facility to the project. We envision that the various (sub)graph matching queries will be specified by constructing a query graph, which will be decorated (via pop-up menus) with selection predicates on the nodes. A palette of graph components (various types of nodes and edges), and panel(s), perhaps in the form of nested directory trees, will be provided for selection of node labels from various taxonomies (bioprocesses, biochemical entities, etc.).

Aim 3. Develop efficient data organization and processing of microarray databases by setting up a database using a standard MIAME scheme (9/03). Work continues on schema based on the MIAME schema with Tony Martino, SNL, and Jerry Timlin, SNL, of Subproject 1. All the web interfaces and data browsing interfaces are generated automatically from the schema by the LBNL-Object-Database-Tools (`LODAT`) we are using. The underlying system is Oracle.

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). We are developing statistical analysis and visualization tools for prediction of functional associations between proteins. We have also begun implementing various computational algorithms to infer functional association links in the protein-protein interaction networks (see Aim 1) reported elsewhere or being developed in-house.
We have also obtained a license for the Columbia University `GRASP` molecular visualization and analysis program (http://trantor.bioc.columbia.edu/grasp/), and have obtained an existing SGI IRIX machine to install this software and have it set up for use by our project members.

Aim 5. Set up environment that allows researchers in this proposal to utilize the computational resources at ORNL and Sandia (9/03). A notarization framework was designed, developed and documented this quarter to support a bio-aware notebook and an interactive work environment. The notarization framework provides strong cryptographic tamper resistance for all the data and notes submitted into the notebook. The notarization also has the capability to provide legally binding time stamps on the submitted entry. The notarization framework provides a key set of basic services for use by the bio-aware notebook, but also can be used by any other GTL applications requiring these services. A user manual for the notarization framework was written and will be placed in the project notebook.

Progress continues to be made in the development of a Matlab-like bioinformatics work environment – named 'BiLab'. A language has been designed that incorporates the scripting style of Matlab with a range of modern language features for larger scale and library style generic and object-oriented software development. In order to leverage both the existing software and the researcher knowledge base, much of the Matlab low-level syntax has been adopted.

Also this quarter we set up (at ORNL) the Sun iForce cluster that we received from NCSA. This cluster contains two Netra T1 150 application servers, and four Enterprise 220R dual cpu machines set up in two clusters of two units. Each cluster shares a D1000 disk farm. One cluster is a high-performance Web Server, and the other an Oracle database server.

Collaboration With Others

Nagiza Samatova, ORNL, (Subproject 5) has initiated a collaboration with Adam Arkin from LBNL on integrating his GTL project with our efforts to develop a protein interaction network browser. She also began a collaboration with Arie Shoshani, LBNL (Subproject 5) and Frank Olken, LBNL (Subproject 5) on providing search, query and retrieval capabilities to the browser. It was agreed that the tools for inferring functional association links between proteins will be shared between ORNL and Adam Arkins's, LBNL, project.

Arie Shoshani, LBNL, (Subproject 5) has been working jointly with Tony Martino (Subproject 1) to develop a web-based Data Entry and Browsing (DEB) tool that will facilitate capturing the metadata from experiments in a computer searchable form. The system is built on top of the Object-Based Database Tools (developed previously at LBNL) and the data is backed up in the Oracle database system. The design of this tool was developed in collaboration with Tony providing provided insight as to the features that a biologist will find useful. One of the salient features of the design is the ability to browse through previous experiments and describe the next experiments with minimal effort based on existing entries in the database. Another important feature is providing security to the object-instance level by users and groups, where each experiment or related objects can be assigned read, write, and delete permission to other users/groups. The programmer who developed the current version is Victor Havin. A non-secure example can be viewed at:
http://sdm.lbl.gov/~opm7/sdmdev/www/jopmDocs/sandia/v2.1/sandiaForm.html.

Gong-Xin Yu and Nagiza Samatova of ORNL (Subproject 5) continued working with Brian Palenik, UCSD (Subproject 1) to identify a set of key genes contributing to marine vs. non-marine discrimination of genomes. A set of target genomes under study have been identified by Brian and initially annotated at ORNL. The application of ORNL tools to this problem will continue in the next quarter.

Al Geist, ORNL (Subproject 5) initiated a dialog with Nancy Slater, LBNL, from the LBNL GTL project on the use of common electronic notebook software and LIMS in the projects. Al has provided the

notebook software used by our GTL project and will follow-up with information on the LIMS system used by the ORNL/PNNL GTL project (PI Michelle Buchanan, ORNL).

Frank Olken, LBNL (Subproject 5) has begun working with Jean Loup Faulon, SNL (Subproject 4) on extended degree sequences for graph characterization. Faulon has speculated that extended degree sequence signatures can be used to aid graph isomorphism queries.

Publications and Presentations

1. Park, H., Ostrouchov, G.,Yu, G.X., Geist, A., Gorin, A. and Samatova, N., "Inference of Protein-Protein Interactions by Unlikely Profile Pair," *Proceedings of the Third IEEE International Conference on Data Mining*, Melbourne, Florida, USA, November 19-22, 2003.

2. Samatova, N., Yu, G., Park, H., Geist, A., Ostrouchov, G., "From Genomics to Functional Proteomics: *In silico* Approach," *A mini-symposium on Parallel Computational Biology* (in conjunction with SIAM Conference on Parallel Processing for Scientific Computing, San Francisco, CA, Invited, February 25-27, 2004.

3. Yu, G.-X., Geist, A., Ostrouchov, G., Samatova, N. F., "An SVM-based Algorithm for Identification of Photosynthesis-Specific Genomes Features," *The IEEE Bioinformatics Conference*, Stanford, CA, August 11-14, 2003.

4. Yu, G.X, Geist, A., Ostrouchov, G., and Samatova, N., "An SVM-based Algorithm for Identification of Photosynthesis-Specific Genomes Features," *Proceedings of the IEEE Bioinformatics Conference*, pp. 235-243, Stanford, CA, August 11-14, 2003.

5. Yu, G.X., Park, H., Chandramohan, P., Munavalli, R., Ostrouchov, G., Geist, A., Samatova, N.F. "A Surface Patch Ranking Method Identifies Correlated Substrate Specificity Residues in Highly Homologous Enzymes," *11th International Conference on Microbial Genomes*, Durham, NC, Abstract, Sept. 28 – Oct. 2, 2003.