OMICS A Journal of Integrative Biology
Volume 6, Number 4, 2002

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

GRANT S. HEFFELFINGER,[1] ANTHONY MARTINO,[2] ANDREY GORIN,[3] YING XU,[3] MARK D. RINTOUL III,[1] AL GEIST,[3] HASHIM M. AL-HASHIMI,[8] GEORGE S. DAVIDSON,[1] JEAN LOUP FAULON,[1] LAURIE J. FRINK,[1] DAVID M. HAALAND,[1] WILLIAM E. HART,[1] ERIK JAKOBSSON,[7] TODD LANE,[2] MING LI,[9] PHIL LOCASCIO,[3] FRANK OLKEN,[4] VICTOR OLMAN,[3] BRIAN PALENIK,[6] STEVEN J. PLIMPTON,[1] DIANA C. ROE,[2] NAGIZA F. SAMATOVA,[3] MANESH SHAH,[3] ARIE SHOSHONI,[4] CHARLIE E.M. STRAUSS,[5] EDWARD V.THOMAS,[1] JERILYN A. TIMLIN,[1] and DONG XU[3]

ABSTRACT

The U.S. Department of Energy recently announced the first five grants for the Genomes to Life (GTL) Program. The goal of this program is to "achieve the most far-reaching of all biological goals: a fundamental, comprehensive, and systematic understanding of life." While more information about the program can be found at the GTL website (www.doegenomestolife.org), this paper provides an overview of one of the five GTL projects funded, "Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling." This project is a combined experimental and computational effort emphasizing developing, prototyping, and applying new computational tools and methods to ellucidate the biochemical mechanisms of the carbon sequestration of *Synechococcus* Sp., an abundant marine cyanobacteria known to play an important role in the global carbon cycle. 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. The project includes five subprojects:

[1]Sandia National Laboratories, Albuquerque, New Mexico.
[2]Sandia National Laboratories, Livermore, California.
[3]Oak Ridge National Laboratory, Oak Ridge, Tennessee.
[4]Lawrence Berkeley National Laboratory, Berkeley, California.
[5]Los Alamos National Laboratory, Los Alamos, New Mexico.
[6]University of California, San Diego, California.
[7]University of Illinois, Urbana/Champaign, Illinois.
[8]University of Michigan, Ann Arbor, Michigan.

an experimental investigation, three computational biology efforts, and a fifth which deals with addressing computational infrastructure challenges of relevance to this project and the Genomes to Life program as a whole. Our experimental effort is designed to provide biology and data to drive the computational efforts and includes 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. Our computational efforts include coupling molecular simulation methods with knowledge discovery from diverse biological data sets for high-throughput discovery and characterization of protein-protein complexes and developing 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. Furthermore, given that the ultimate goal of this effort is to develop a systems-level of understanding of how the *Synechococcus* genome affects carbon fixation at the global scale, we will develop and apply a set of tools for capturing the carbon fixation behavior of complex of *Synechococcus* at different levels of resolution. Finally, because 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, we have also established a companion computational infrastructure to support this effort as well as the Genomes to Life program as a whole.

INTRODUCTION

As described in the Genomes to Life (GTL) Program literature (www.doegenomestolife.org), the goal of this program is to "achieve the most far-reaching of all biological goals: a fundamental, comprehensive, and systematic understanding of life":

> DOE's Genomes to Life program will make important contributions in the quest to venture beyond characterizing such individual life components as genes and other DNA sequences toward a more comprehensive, integrated view of biology at a whole-systems level. The DOE offices of Biological and Environmental Research and Advanced Scientific Computing Research have formed a strategic alliance to meet this grand challenge.
>
> The plan for the 10-year program is to use DNA sequences from microbes and higher organisms, including humans, as starting points for systematically tackling questions about the essential processes of living systems. Advanced technological and computational resources will help to identify and understand the underlying mechanisms that enable organisms to develop, survive, carry out their normal functions, and reproduce under myriad environmental conditions.
>
> This approach ultimately will foster an integrated and predictive understanding of biological systems and offer insights into how both microbial and human cells respond to environmental changes. The applications of this level of knowledge will be extraordinary and will help DOE fulfill its broad missions in energy, environmental remediation, and the protection of human health.

The Genomes to Life program has four stated goals:

1. Identify and characterize the molecular machines of life—the multiprotein complexes that execute cellular functions and govern cell form

2. Characterize gene regulatory networks
3. Characterize the functional repertoire of complex microbial communities in their natural enviroments at the molecular level
4. Develop the computational methods and capabilities to advance understanding of complex biological systems and predict their behavior

As stated above, the effort discussed in this paper is focused on understanding the carbon sequestration behavior of *Synechococcus* Sp. through experimental and computational methods with the major effort in the development of computational methods and capabilities for application to *Synechococcus*. Thus, this project can be thought of as a "Goal 4" project. To this end, the work has been divided into five subprojects:

1. Experimental Elucidation of Molecular Machines and Regulatory Networks in S*ynechococcus* Sp.
2. Computational Discovery and Functional Characterization of *Synechococcus* Sp. Molecular Machines
3. Computational Methods Towards the Genome-Scale Characterization of *Synechococcus* Sp. Regulatory Pathways
4. Systems Biology for *Synechococcus* Sp.
5. Computational Biology Work Environments and Infrastructure

These five subprojects are discussed individually in the sections that follow, 1.0 through 5.0, respectively.

The computational work in this proposal is captured in sections 2.0, 3.0, 4.0, and 5.0, while the experimental biology, including experimental methods development, required to integrate and drive the computational methods development and application are discussed in 1.0. "Computational Discovery and Functional Characterization of *Synechococcus* Sp. Molecular Machines," discussed in section 2.0, is aimed directly at the GTL Goal 1 in the context of *Synechococcus* and includes primarily computational molecular biophysics and biochemistry as well as bioinformatics. "Computational Methods Towards the Genome-Scale Characterization of *Synechococcus* Sp. Regulatory Pathways," section 3.0, is also highly computational, focused primarily on the development and application of bioinformatics and data mining methods to elucidate and understand the regulatory networks of *Synechococcus* Sp. (GTL Goal 2).

In section 4.0, we discuss our planned efforts to integrate the efforts discussed in sections 1.0, 2.0, and 3.0 to enable a systems biology understanding of *Synechococcus*. This work will support GTL Goals 1 and 2 and is focused on developing the computational methods and capabilitities to advance understanding of *Synechococcus* as a complex biological system. Given the available information and data on *Synechococcus,* the effort discussed in section 4.0 will initially (year 1) employ other microbial data in order to advance the state of the art of computational systems biology for microorganisms. This will give our *Synechococcus* experimental effort (section 1.0) time to ramp up and produce the data needed to drive this effort in project years 2, and 3 (FY04 and FY05).

In section 5.0, "Computational Biology Work Environments and Infrastructure," we discuss a number of developments to enable the support of high-throughput experimental biology and systems biology for *Synechococcus* including work environments and problem solving environments, as well as high performance computational resources to support the data and modeling needs of GTL researchers.

Synechococcus

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 environmental response to anthropogenic inputs of CO_2 into the atmosphere and in global carbon modeling. However, the actual biochemical mechanisms of carbon fixation and their genomic basis are poorly understood for these organisms as is their relationship to important macroscopic phenomena. For example, we still do not know what limits carbon fixation in many areas of the oceans. Linking an organism's physiology to its genetics is essential to understand the macroscopic implications of an organism's genome (i.e., linking "genomes to life").

The availability of S*ynechococcus'* complete genome allows such an effort to proceed for this organism. Thus the major biological objective of this work is to elucidate the relationship of the *Synechococcus'* genome to *Synechococcus'* relevance to global carbon fixation through careful studies at various length scales and levels of complexity. To this end, we will develop a fundamental understanding of binding, protein complexes, and protein expression regulation in order to provide a complete view of the protein binding domains that mediate the most relevant protein–protein interactions and their related regulatory networks. In addition, we will investigate molecular machines and regulatory networks within the *Synechococcus* cell. Specifically, we will develop a fundamental understanding of the protein binding domains that mediate protein–protein interactions and form the basis of the *Synechococcus* molecular machines most relevant to carbon fixation. In addition, we will investigate *Synechococcus'* regulatory network and study a few mo-lecular machine complexes in detail. Our goal will be to elucidate the fundamental information regarding binding, protein complexes, and protein expression regulation to enable a systems-level understanding of carbon fixation in *Synechococcus.*

1.0 EXPERIMENTAL ELUCIDATION OF MOLECULAR MACHINES AND REGULATORY NETWORKS IN *SYNECHOCOCCUS* SP.

The subproject will focus on three protein families: the carboxysome, ATP-binding cassette (ABC) transporters, and histidine kinase-response regulators. The carboxysome is a complex of proteins that converts inorganic carbon (CO_2) to sugars via the photosynthetic carbon reduction chain. It is the primary organelle involved in carbon fixation, yet its structure and function are not well understood. ABC transporters and histidine kinase-response regulators appear to be intricately linked. The kinase-regulator signal transduction system detects environmental changes and regulates cellular response by inducing appropriate genes. Cellular response often includes the influx and efflux of carbon and nutrients; thus, the kinase-regulators affect ABC transporters, proteins involved in active transport. The mechanisms by which these two systems are co-regulated are not understood.

In the text that follows, we outline some of the important characteristics and questions regarding carboxysomes, ABC transporters, and histidine kinase-response regulators, and discuss how we will analyze these interdependent protein complexes using mass spectrometry, display techniques, and gene microarray analysis. Finally, we discuss our plans for employing our experimental program to integrate and prototype the computational capabilities discussed in the sections that follow.

Carboxysomes and inorganic carbon fixation

Carboxysomes are polyhedral-shaped inclusion bodies approximately 100 nm in diameter and present in cyanobacteria and many chemoautotrophs (Cannon et al., 2001). They consist of a protein shell and are filled with the enzyme ribulose 1,5 bisphosphate carboxylase/oxygenase (RuBisCO). Carboxysomes maintain two known functions. They act in a carbon-concentrating mechanism (CCM) to increase the concentration of CO_2 from the relatively low value of 20 μM in seawater, and they convert CO_2 to sugars. In fact, RuBisCO catalyzes the photosynthetic conversion.

The functions of the carboxysome are dependent on the protein composition and structure of the complex. In *H. neapolitanus,* SDS-PAGE analysis indicates the carboxysome consists of nine to fifteen polypeptides of which at least six are associated with the shell (Cannon et al., 1983; Holthuijzen et al., 1986). In the freshwater strain *Synechococcus* PCC7942, up to 30 different polypeptides are evident, but of the 30 only 10 have similar molecular weights as those found in other organisms (Price et al., 1992). It is likely that many polypeptides represent contaminants, as protein purifications are difficult in cyanobacterium due to high membrane content. Genetic analysis has provided some information as to the likely nature of the shell composition. In *H. neopolitanus, csoS1* encodes a shell protein that has been duplicated twice (*csoS1B* and *csoS1C*) (English et al., 1994) and is highly homologous to the *Synechococcus* PCC7942 genes *ccmK* and *ccmO* (Shively et al., 1996). *CsoS1* and *ccmK* and *-O* genes exist in *cso* (carboxysome) and *ccm* (carbon-concentrating mechanisms) gene clusters that, in their respective organisms, are proximal to the

genes that code for RuBisCO large and small subunits. Functions for most of the genes within the clusters, though, remain unknown.

The protein composition of carboxysomes is especially important in their role in the CCM. The CCM provides concentrated CO_2, the substrate for RuBisCO, through dehydration of HCO_3^-, a relatively high concentration (2 mM) source of inorganic carbon in seawater. HCO_3^- is converted to CO_2 by the enzymatic activity of carbonic anhydrase (CA). It is not clear how high a concentration CO_2 is made available to RuBisCO. It is possible that CA is present in the carboxysome. It has been suggested that CA is sequestered with RuBisCO in the carboxysome in some systems (Fridlyand et al., 1996; Kaplan et al., 1989), but not in others (Bedu et al., 1992; Lanaras et al., 1985). To date, a direct biochemical link between CA and the carboxysome in the CCM has not been established. Additionally, CA genes have not been found to map to carboxysome genome regions.

A number of questions with regard to carboxysome structure and function remain open. The number of proteins, the composition of the shell, and the localization of proteins encoded by functionally important gene clusters are not known. It appears that more proteins are present than can be accounted for by RuBisCO and the shell, and it is possible that CA is present acting independently or in synergy with other components. Finally, molecular interactions between protein components are not known.

The ABC transporter and histidine kinase-response regulator systems

ABC transporters are a superfamily of proteins that transport a wide variety of solutes including amino acids, ions, sugars, and polysaccharides. These transporters have four domains, two hydrophobic integral membrane protein domains and two hydrophilic ATP-binding domains that are thought to couple ATP hydrolysis to transport (Chang et al., 2001). The domains can be found as separate proteins or as a single fused protein. In bacterial systems involved in solute uptake, there is a separate solute binding protein that binds the solute being transported, docks with the other components, and allows the solute to diffuse through a channel before disassociating. Clearly, ABC transporters are sophisticated protein machines that carefully recognize particular compounds and move them into and out of the cell. In *Synechococcus* WH8102, there are about 80 genes that encode for ABC transporters, including about eighteen specific to substrate-binding proteins.

The regulation of transport is a complex multi-level process. The two-component histidine kinase-response regulator signal transduction system appears to play a critical role (Hoch et al., 1995) in this process. Histidine kinase proteins or protein domains sense some property and activate or repress through phosphorylation a second protein called a response regulator. The response regulator acts as a transcription factor and regulates gene induction levels. For instance, sensing increased phosphate levels could lead through the two-component system to increased induction of genes that code for affinity phosphate transporter or binding proteins. In fact, two component regulatory systems have been linked to phosphate transport, nitrogen transport, and porin regulation (Ninfa et al., 1995; Pratt et al., 1995). In *Synechococcus* WH8102, there are six histidine kinases and nine response regulators, and a number of these genes are located physically adjacent to transporters. The genetic information supports biochemical links between regulatory and transport systems.

There are a number of questions pertinent to the level of interaction between the signal transduction components and ABC transporters. It is interesting to consider what histidine kinases and response regulators affect specific ABC transporters and what is the level of regulation in transporter function. Also to be considered is whether multiple transporter functions are effected concurrently and whether there is cross talk between signals.

Protein complex analysis by mass spectrometry

Protein complex compositions and structures, particularly in the carboxysome, will be explored using a couple of techniques that make use of protein mass spectrometry. Novel methods to analyze proteome-scale protein complexes use an affinity purification technique combined with protein identification by mass spectrometry (Gavin et al, 2002; Ho et al., 2002). Cassettes containing tags (poly-His or Protein A) will be inserted at the 3′ end of the genes encoding proteins central to the carboxysome in *Synechococcus.* After clonal selection, cultures will be grown under various conditions of light flux, CO_2 levels, and nutrient concentrations. Cellular lysates will be subjected to tandem affinity purification utilizing low-pressure columns

in order to "fish out" bait proteins and proteins complexed to them (Puig et al., 2001). Proteins will be eluted off the column and separated by SDS-PAGE, and individual protein bands will be excised, digested, and subjected to mass spectrometer analysis. We will conduct protein mass spectrometry with a Bruker Apex FTICR instrument with an electrospray ionization source and a Thermoquest LCQ ion trap interfaced with micro-HPLC. This strategy has the advantages that it is adaptable to high-throughput techniques and does not require the use of antibodies for protein identification (a situation where *a priori* information as to the identity of the protein is needed).

Mass spectrometry will also be used to isolate protein-protein interactions. Protein complexes will be collected as before with tandem affinity purification. In a new step, the protein complexes bound to the column will be chemically crosslinked with amine- or sulfhydryl-specific crosslinkers. The chemically crosslinked complexes will be digested with trypsin, and the peptides separated by capillary reversed-phase HPLC and analyzed by FTICR-MS. Comparing crosslinked data sets to individual protein analysis provides shifts in mass spectra that can be compared to theoretical distance constraints to calculate three dimensional protein arrangements and likely protein–protein interactions.

Protein complexes and networks by phage display

In addition to mass spectrometry, protein complexes can be characterized by using phage display techniques to identify protein-binding domains. Phage display techniques refer to methods by which random samplings of peptide fragments are presented on the surface of viral particles (phage) and are screened for peptide-protein interactions with a probe protein (Smith et al., 1997). A library of degenerate oligonucleotide inserts is cloned into the coding sequence of phage coat proteins so that each viral-containing clone displays a peptide on its surface corresponding to a specific sequence within the library. A probe protein is fixed to any number of possible substrates, and peptide-protein interactions are elucidated by mixing library-containing clones with the probe, selecting and amplifying positives, and repeating the process 2–3 times in a process called panning. Advantages of phage display include the ability to immediately isolate molecular recognition domains and ligand sequences, the ability to construct different libraries to study particular problems, and the ability to screen up to 10^{10} different peptides at once.

We will use phage display using key carboxysome constituents as probe proteins. Libraries of random peptides and libraries based on gene clusters known to be associated with carboxysomes will provide two techniques to establish protein-peptide interactions. In the case of random peptide libraries, the genome will be searched for regions that code for positives indicating potentially new protein-protein interaction pairs. Peptide libraries developed from carboxysome gene clusters represent targeted efforts. Cognate binding partners will be verified by enzyme-linked immunosorbent assay (ELISA) or by yeast 2-hybrid screening.

In addition to protein complex characterization, phage display is a powerful technique that has been used to predict entire protein networks (Tong et al., 2002). As protein binding domains mediate protein-protein interactions, characterizing binding domains can be used to infer networks by developing lists of probable protein interactions. Similar strategies will be employed here. Binding domains are defined in families according to similarities in sequence, structure, and binding ligands. Each member of a family binds a similar sequence, but binding for any particular family member is specified by variances within the core binding domain. Using phage display, we will study known prokaryotic binding domain families in an effort to develop recognition rules. Protein binding domains known to exist in *Synechococcus* include SH3, PDZ, CBS, and TPR domains (Falzone et al., 1994; Bateman et al., 2002).

Gene and protein regulatory networks by microarray analysis

As discussed, the WH8102 genome has six histidine kinases and nine response regulators that are major components in its regulatory network. They are thought to control the sensing of the major nutrients of phosphate, nitrogen, and light. The immediate objective of our experimental investigations is to define the web of interactions controlled by these components with the ultimate goal to developing a systems-level understanding of *Synechococcus.* Our efforts will focus on characterizing the regulation of the transport genes, two component systems, and some stress-associated genes using a DNA microarray containing the entire genome. We will use a gene knockout strategy whereby all histidine kinases and many of the re-

sponse regulators will be inactivated in turn, and the effect of the gene knockouts on the induction of ABC transporters will be measured as a function of nutrient and light levels. Microarray analysis will be completed using state-of-the-art hyperspectral scanning. Using bioinformatic analyses to characterize the upstream regions of the genes regulated by a particular stress, common cis-regulatory sites potentially used by the response regulators will be predicted. Then, the complete genome can be searched for other putative sites with these sequences (see below). In all cases, binding to DNA will be verified experimentally.

One of the advantages of *Synechococcus* as a model system is that bioinformatic analyses can incorporate the data for the complete genomes of the related cyanobacteria *Prochlorococcus* in both the sequence definition phase and the sequence scanning phase. For example, if a sequence motif is found upstream of a gene in all three genomes during genome scanning, a logical prediction is that these genes are regulated in similar ways.

ABC transporter function and regulation will also be analyzed by studying protein expression levels. We will follow the predicted eighteen substrate binding proteins involved in the ABC transporter system using polyclonal antibodies to each protein. Antibodies will be produced in chickens and rabbits from proteins purified from *E. coli* expression of PCR amplified gene segments. Protein expression will be analyzed by SDS-PAGE and Western blotting. Eventually, protein arrays will be developed. Whole cells will be labeled with an amine reactive fluorophore. The fluorophore migrates into the periplasmic space (but not inside the cell) and will label the lysines of proteins in this region and the cell surface. We will then solubilize the cells and incubate the labeled proteins with the array of antibodies and scan the membrane to quantity the amount of each protein. As the signal should be proportional to the number of accessible lysine-residues on each protein, we will carefully calibrate the signal from each protein against known amounts of protein obtained from the expressed protein in *E. coli.*

The experimental design with respect to computational biology

The comprehensive experimental program is designed to integrate and prototype a number of computational capabilities with the goal of developing iterative and synergistic efforts. Whole protein complex characterization supports Rosetta-like efforts to analyze protein structure, and breaking the complexes down into the fundamental binding units supports molecular biophysical calculations. Furthermore, experimental and theoretical results describing binding domains and recognition rules are critical in developing predicted protein interaction maps. The analysis of gene and protein regulatory networks supports a number of efforts from the acquisition, storage, and analysis of microarray data to the development of supervised and unsupervised clustering techniques to predict cis-regulatory sites, regulatory networks, and protein function. These and other computational efforts are outlined below.

2.0 COMPUTATIONAL DISCOVERY AND FUNCTIONAL CHARACTERIZATION OF *SYNECHOCOCCUS* MOLECULAR MACHINES

Our computational discovery and characterization of molecular machines effort is designed to develop high-performance computational tools for high-throughput discovery and characterization of protein–protein complexes. This will be achieved primarily through coupling of molecular simulation methods with knowledge discovery from diverse biological data sets which can then be applied, in conjunction with experimental data, to the *Synechococcus* proteome to enable discovery and functional annotation of protein complexes. Our efforts will be pursued through several synergetic approaches: low-resolution high-throughput Rosetta-type algorithms, high performance all-atom modeling tools, and knowledge-based algorithms for functional characterization and prediction of the recognition motifs. These are discussed individually in the text that follows.

Rosetta-type algorithms

There are currently no highly reliable tools for modeling of protein-protein complexes. Building upon proven methods for ab initio protein modeling, we will develop and apply Rosetta-like algorithms for fast

characterization of protein-protein complexes complexes with two approaches: (1) for cases where structures of unbound members are known, the Rosetta potential will be used to dock them together while permitting conformational changes of the components, and (2) if experimental data are available, sparse constraints will be incorporated (from NMR and mass-spectroscopy experiments). Both approaches will help achieve the goal of developing high-throughput methods of characterizing protein–protein complexes.

All-atom simulations

Our existing parallel codes for biomolecular-scale modeling will be extended as necessary to model protein–protein complexes in *Synechococcus.* All-atom simulations will be initially focused on two problems: (1) interpretation of the phage display data and (2) investigation of the functional properties of *Synechococcus* membrane transporters. The computational algorithms and software developed in this effort will be applicable broadly to molecular machines in other organisms to provide understanding of protein interactions in general.

Knowledge fusion algorithms

Because existing data mining algorithms for identification and characterization of protein complexes are not sufficiently accurate, nor do they scale well for genome-wide studies, we will extend or develop new algorithms to improve predictive strength and allow new types of predictions to be made. Our approach will involve (1) developing "knowledge fusion" algorithms that combine many sources of experimental, genomic and structural information, (2) coupling these algorithms with modeling and simulation methods, (3) implementing high performance, optimized versions of our algorithms. Specifically algorithms for three interrelated problems will be investigated: (1) identification of pair-wise protein interactions, (2) construction of protein-protein interaction maps, and (3) functional characterization of the identified complexes.

Together these three elements yield a synergistic computational approach to discovering and characterizing molecular machines. Thus, for example, protein pair identification tools will be used to provide the initial sets of putative pairs of interacting proteins, either by filtering our experimental data or bioinformatics leads (from efforts described in section 3.0) for specific metabolic subsystems of *Synechococcus.* This initial set of targets and the available experimental constraints will be investigated further through the use of the Rosetta-like algorithms and all-atom methods. The resulting information will then be used to refine the knowledge fusion algorithms as well as applied for the functional characterization of the verified protein assemblies.

Background

Genome-scale techniques for measuring, detecting, mining, and simulating protein-protein interactions will be critical for transforming the wealth of information currently being generated about individual gene products into a comprehensive understanding of the complex processes underlying cell physiology. Current approaches for accomplishing this formidable task include direct experimentation, genome mining, and computational modeling. This effort will exploit all three approaches. Below we briefly discuss the current state-of-art and existing limitations of these approaches.

The leading experimental genome-wide high-throughput methods for characterizing protein-protein interations include the two-hybrid system (Fields et al., 1989; Uetz et al., 2000), protein arrays (Finley et al., 1994), and the phage display (Rodi et al., 1999). Although direct identification methods provide wide genome coverage, they have a number of limitations intrinsic to their experimental design. First, a protein must preserve a correct fold while attached to the chip surface (or linked to the hybrid domain). Otherwise, the method can capture non-native interactions. Second, the binary nature of these approaches is even more restrictive because many of the cellular machines are multiprotein complexes, which may not be fully characterized by pair wise interactions. Finally, short-living protein complexes are a tremendous problem for all of these methods. Transient protein–protein complexes are thought to comprise a significant fraction of all regulatory interactions in the cell and may need additional stabilization for the detection.

Over the last 5 years, experimental approaches have been supplemented by bioinformatics methods based on genome context information. These methods explore correlations of various types of gene contexts and functional interactions between corresponding encoded proteins. Several types of genomic context have been utilized including: fusion of genes (Marcotte et al., 1999; Enright et al., 1999), also called the Rosetta Stone approach; co-occurrence of genes in potential operons (Overbeek et al., 2000; Dandekar et al., 1998), and co-occurrence of genes across genomes (Pellegrini et al., 1999) which is based on an assumption that proteins having similar phylogenetic profiles (strings that encode the presence or absence of a protein in every known genome) tend to be functionally linked or to operate together.

Unfortunately the usefulness of many valuable bioinformatics methods are seriously limited due to (1) high loads in false negatives (resulting from incomplete coverage) and false positives (resulting from indirect interference detection), (2) low genome coverage due to a low percentage of genes that meet underlying assumptions (e.g., in a comparative study by Huynen (Huynen et al., 2000), the conservation of gene order for Mycoplasma genitalium had the highest coverage, 37%, among all available genomes and all considered methods), and (3) predictions that are mostly derived from sequence analysis and do not incorporate any information about the structure of the interacting proteins.

For these reasons, the full power of bioinformatics approaches realized only in close integration with experimental and/or other computational methods. We will use such a collaborative approach in this effort as we develop new identification algorithms, combining information from several heterogeneous sources.

Development of rosetta-based computational methods for characterization of protein–protein complexes

The computer program "ROSETTA" is currently the leading program for protein structure prediction (rated first in CASP-2001), and its approach represents a powerful foundation for applying computational methods for characterization of protein-protein complexes. In this work, we will create a tool that will enable the assessment of the probability of candidate pairs of proteins forming a complex, assuming known structures. This tool will be immensely useful for many applications aimed at genome-level categorization. One example is the prediction of putative binding partners given a set of proteins with known structures. As the number of known protein structures grows exponentially, such a question will soon embrace a significant part of the bacterial proteome. Ultimately, we hope to obtain a detailed characterization of the obtained molecular machine (e.g., footprint interaction areas and spatial parameters of the molecular machines) leading to accelerated functional characterization and a better prediction of the recognition motifs. Below we describe our research strategy in detail: how we will tune the program to protein complex characterization, our plans for integrating experimental constraints, and several possible strategies for increasing Rosetta performance.

In the past, Rosetta has mainly been applied to the task of predicting the entire, end-to-end, structures of protein domains starting from just the primary amino acid sequence. The main limitation of this ab initio protein structure prediction approach has been protein size, given the computational difficulties of the method. In this work, we will employ the Rosetta method in a different fashion. Starting from a set of proteins whose unbound structure is mostly known from experimental measurements, but whose arrangement in a complex is the subject of interest, we will tune the Rosetta algorithm to the protein complex characterization problem. This will require several innovations of the method: (1) introducing a probability distribution for the distance between two centers of mass for individual chains, (2) adapting the Rosetta potential functions to identify the optimal docking arrangement between two proteins in a complex, and (3) adapting Rosetta to allow for conformational changes induced by the binding event.

Two primary questions will be addressed: (1) which proteins are binding partners and (2) how, physically, do they bind? The first question can be answered by trying all combinations of putative partners in a set and assessing the relative quality of the fit or misfit for the binding configurations predicted by Rosetta; from this a probability of binding can be estimated. The second question can be addressed by reporting the optimal configuration found by Rosetta for any given pair. In practice, one would report not a single conformation but rather multiple plausible configurations, and then corroborate these by attempting to refine

the interaction configuration by molecular dynamic (MD) simulation and other all-atom modeling methods in a high throughput implementation.

We will develop our technique in stages, moving from small, static protein–protein models to large complex ones, in the following timeline:

1. Simulate the protein–protein complex available from PDB
2. Simulate the subdomain-peptide recognition, comparing with data from phage display experimental information
3. Study the assembly of small ribosomal proteins where RDC constraints will guide the assembly procedure
4. Carry out simulations without using experimental constraints
5. Conduct studies on a large set of proteins from *Synechococcus* identifying targets with tools developed in cooperation the experimental group and the bioinformatics group

We will progress to systems involving multiprotein complexes, protein–protein complexes involving dynamic behavior of the components, and systems with a large number of decoys (proteins that are proven not to be putative binding partners). It is anticipated that some of the most interesting biological systems in the foreseeable future will be reachable only for "low resolution" technologies such as Rosetta. For one example, we may consider multiprotein complexes involving 40–60 proteins (e.g., cellular membrane assemblies).

Mass spectroscopy coupled with cross-linking (MS/CL) and NMR experiments are very attractive technologies for characterizing the molecular machinery in the cell. Integration of such experimental information with the Rosetta method is a centerpiece of our strategy, as constraint-supported docking is likely to have a much higher success rate for correct identification of the complexes. This experimentally obtained information will be used to train our knowledge-based prediction methods tools, improve the Rosetta potentials and ultimately enable us to reduce our dependence on constraints and successfully perform unconstrained simulations/docking of selected complexes.

High-performance all-atom modeling of protein machines

We will model two "molecular machine" problems in *Synechococcus*. In the first effort, we interpret data from phage display experiments (section 1.0); in the second we will deduce functional properties of *Synechococcus* membrane transporters.

The phage display library screens discussed in section 1.0 for *Synechococcus* proteins will yield ligands that bind to specific proteins. Due to uncertainties (e.g., counts of expressed ligands on phage surfaces, alteration in binding strength due to ligand tethering, calibration of fluorescence measurements, etc), these experiments will provide only a qualitative measure of binding affinity. Thus the relative binding strength of an individual ligand/protein pair cannot be accurately compared to other pairings. We will use molecular-scale calculations to compute relative rankings of affinities for the ligands found to bind to each probe protein in the phage library screens. These rankings will be used in the protein/protein interaction models discussed in section 4.0. Additionally, we will identify mutated ligand sequences with likely binding affinity that can be searched for within the *Synechococcus* proteome to infer protein/protein pairings beyond those indicated by the phage experiments. This work will proceed in 2 stages: we will first compute ligand conformations, then perform flexible docking of ligands to the known binding domains of the target proteins.

In phage display experiments a short peptide chain (ligand) is expressed on a phage surface where it potentially binds to a protein (the probe or target) in the surrounding solution. The ligand is fused to coat proteins (typically pVIII or pIII proteins) on the phage surface. We will model ligand conformation and orientation (relative to the phage surface) for representative ligands found as hits in the library scans performed experimentally, and thus inferred to bind to specific prokaryotic protein motifs in *Synechococcus*. Because the ligands are short (9-mers to 20-mers), we anticipate being able to compute their structure "de novo," using a combination of computational approaches: Monte Carlo, molecular dynamics and parallel tempering. In all of these methods, water can be explicitly treated, which is a critical contributor to the native structure of the ligand in an aqueous solution. The tethering of the ligand to the phage surface can also be

naturally included in the models, as can the presence of the phage surface, which affects the energetics of the ligand conformation and the ligand/water interactions.

Atomistic Monte Carlo (MC) techniques will first be used for this problem. Sandia's Towhee MC code (Martin et al., 2000) can perform simple MC moves as well as protein-specific moves (e.g., torsional rotations of backbone bonds, side-chain regrowths) to sample a peptide chain's conformational space. Towhee also has a configurational-bias (CB) MC capability (Martin et al., 1999) so that a peptide chain can be "grown," atom by atom, into a surrounding medium in a manner that preferentially generates lower energy conformations. Towhee can then solvate the system with water using similar CB-MC insertions.

We will also investigate the use of a new method, parallel tempering (or replica-exchange) (Mitsutake et al., 2001), to generate low-energy ligand conformations. In parallel tempering, multiple copies of a molecular-scale simulation are created and simulated at different temperatures using traditional MD. Periodically, the respective temperatures of a pair of ensembles are swapped according to Monte Carlo rules. The method is highly parallel since individual replicas run with little communication between them. Parallel tempering can find low-energy conformations much more quickly than a standard MD simulation. Garcia et al. (2001) used these methods to find the native beta-hairpin conformational state of an 18-mer peptide fragment of protein G in explicit solvent within a few nanoseconds of MD simulation time, starting from a denatured conformation. Similar work predicted alpha-helical structures in short peptide chains (Sanbonmatsu et al., 2002). We propose to enhance our LAMMPS MD code to include a replica-exchange capability whereby $P = M \times N$ processors can run M replicas, each on N processors. This will enable us to efficiently apply all the LAMMPS features (particle-mesh Ewald, rRESPA, force fields) to computing ligand conformations.

All-atom methods will also be applied to investigate ligand conformations for docking. Both high-accuracy conformations and low-accuracy conformations will be investigated to maximize the information yielded given the computational requirements of the flexible docking methods we will employ.

Transport proteins found in cell membranes are known to be important to the functioning of *Synechococcus,* as to all microbes and all cells. Yet much about these molecular machines is unknown, from the function and regulation of individual transporters to the interaction and cross-talk between multiple transporters. We will model three types of transporters in *Synechococcus:* ion channels, small multi-drug resistance (SMR) transporters, and ATP binding cassette (ABC) transporters. The goal of this effort is to uncover the physical basis for the function of these transporters. These studies will provide molecular insight for the system-level cell models developed in this effort (section 1.0), for example, as boundary conditions on the cell as it interacts with its extracellular environment.

"Knowledge fusion" based characterization of biomolecular machines

Developing new data mining and statistical analysis methods for the identification of protein-protein interactions is an urgent need for several reasons. First, interactions can be deduced in unusual ways from many very diverse data sources (for example, from the fact that genes from one genome are fused in genome of another organism). Second, information is accumulating in databases of all kinds (complete genomes, expression arrays, proteomics, structural) at an unprecedented rate, resulting in a landslide of unanalyzed data. Furthermore, taking advantage of the information in these structural and biophysical databases presents a special challenge as many existing data mining techniques were developed for sequence databases and conceptually new approaches will be needed for the structural domain.

The focus of this task is to develop advanced data mining algorithms to elucidate (1) which proteins in a cell interact both directly (via multiprotein complex) and indirectly (via biochemical process, metabolic or regulatory pathway), (2) where on the protein surface the interaction occurs, and (3) what biological functions the protein complex performs. Existing data mining tools for making such inferences have low predictive accuracy and do not scale for genome-wide studies. This is largely due to incorporation of data at a single or very few levels, lack of sufficient data, and/or computational intractability of exact algorithms. We will improve the predictive accuracy of such methods through several approaches: developing "knowledge fusion" based algorithms that make predictions by fusing knowledge extracted from various sources of bioinformatics, simulation, and experimental data, coupling these algorithms with modeling and simu-

lation methods for approximating structure-related missing data, and extending the applicability of these algorithms to the genome-scale through developing their high performance optimized versions suited for Terascale computers.

Our development strategy will involve three parts: identification of pair-wise protein interactions, construction of protein interaction maps of these complexes, and functional characterization of identified complexes. These tools will be prototyped with application to the *Synechococcus* proteome in coordination with our regulatory pathway mining effort and used to obtain information necessary for our systems biology effort.

Discovery and characterization of Synechococcus *molecular machines*

Our aim is to apply the computational methods developed in this effort to *Synechococcus.* We will initially verify known molecular machines in *Synechococcus* to prototype our methods. Ultimately we expect that these methods will enable us to (1) discover novel multiprotein complexes and protein binding domains that mediate the protein-protein interactions in *Synechococcus,* and (2) better understand the functional mechanisms involved in carbon fixation and environmental responses to carbon dioxide levels. In particular we will characterize *Synechococcus* protein-protein interactions that contain known interaction domains such as SH3 domains, and LRRs, as well as the *Synechococcus* protein complexes related to carboxysomal, ABC transporter systems and also protein-protein interactions involved into circadian system and light-signal transduction pathways.

About 10% of the genes of bacterial genomes are dedicated to transport and there are approximately 200 transporter families. Validations and applications of our biomolecular machines characterization pipeline methods will be tested by focusing on elucidating the functional mechanisms of protein complexes related to carboxysomal and ABC transporter systems in *Synechococcus.* The categorical data analysis based prediction methods will be applied to all amino-acid sequences of interest in *Synechococcus* genome. This will generate a probability matrix with a probability of interaction assigned to each protein pair. Rosetta-based modeling methods will be applied to a selected set of more likely interacting protein pairs. This will provide a basis for determining putative structural properties of selected proteins and give hints about potential protein-protein interaction residue sites. The identified structural properties will be used by prediction methods to further validate and/or refine a set of interacting protein pairs. Thus, these knowledge-based prediction methods coupled with modeling and simulation will determine a set of possible protein pairs involved in the carboxysomal and ABC transporter complexes and a set of their putative binding sites.

3.0 COMPUTATIONAL CHARACTERIZATION OF BIOLOGICAL PATHWAYS

In living systems, control of biological function occurs at the cellular and molecular levels. These controls are implemented by the regulation of activities and concentrations of species taking part in biochemical reactions (Stephanopoulos et al., 1998). The complex machinery for transmitting and implementing the regulatory signals is made of a network of interacting proteins, called regulatory networks. Characterization of these regulatory networks or pathways is essential to our understanding of biological functions at both molecular and cellular levels. Traditionally, study of regulatory pathways is done on individual basis through *ad hoc* approaches. With the advent of high-throughput measurement technologies, e.g., microarray chips for gene/protein expression and two-hybrid systems for protein-protein interactions, and bioinformatics, it is now feasible and essential to develop new and effective protocols for tackling the challenge of systematic characterization of regulatory pathways. The impact of these new high-throughput methods can be greatly leveraged by carefully integrating new information with the existing (and evolving) literature on regulatory pathways in all organisms. The main focus of this subproject is to develop a suite of computational capabilities for (1) information extraction from large-scale biological data, including microarray gene expression data, genomic sequence data, protein-protein interaction data, and (2) systematic construction of biological pathways through fully utilizing information extracted from large-scale biological data mining and analysis. These capabilities will improve the state of the art in biological pathway inference, in terms

of both prediction reliability and application scope. Initially, these capabilities will be applied to a set of selected pathways in *Synechococcus,* including ABC transporter pathway. As the technologies mature, they will be ported on to high-performance supercomputers to provide community-wide services for general biological pathway constructions.

Derivation of regulatory pathways through combining multiple sources of information: our vision

Given the complexity of regulatory pathways and incomplete nature of existing high-throughput experimental data, it is unlikely that one could develop a computational capability for accurate and automated derivation of novel biological pathways in the near future. It will be more realistic if we set our goal to identify which parts of a target pathway are inferable from the publicly available data/information and to construct them; and then to identify what data we may need, specific to a particular pathway, for a full characterization of the pathway. Through rational design of experiments for further data collection to fill in the "gaps," we can significantly reduce the cost and time needed to fully characterize a pathway. To make the experimental data more useful, we will first develop a number of improved capabilities for generation and interpretation of data. Initially these data include (a) microarray gene-expression data, (b) genomic sequence data, and (c) protein–protein interaction data. Then we will investigate an inference framework for pathways, which can make use of all of these data. This inference framework will be able to pull together pathway information from our own work and from earlier relevant investigations. With that corpus we will be able to (1) to assign weights to each data item, based on our assessment on the quality of each data source and the cross-validation information from other sources, (2) identify components of a target pathway and their interaction map to the extent possible, and (3) identify the parts of a target pathway that are not inferable from the available data. This framework will be organized such that new sources of information or analysis tools can be easily added as they become available, without affecting the other parts of the framework. We envision that we can quickly generate a set of possible candidate pathway models, possibly with certain parts missing or uncertain, using this inference framework. An iterative process will then follow to design and conduct experiments through rational design and then feed the new and more specific data to this inference framework to refine the models. As more high-throughput genomic/proteomic data become available, we can expect that fewer pathway-specific experiments will be needed for complete characterization of a pathway and hence the more automated this inference framework will become.

Improved experimental and computational capabilities for high-throughput biological data generation and interpretation

High-throughput biological data generation capabilities, like microarray gene expression chips or two hybrid systems for protein-protein interactions, provide powerful tools for scientists to investigate the internal structures and functional mechanisms of various biological processes at the molecular level. However it is well known that these high-throughput technologies, at the current stage, could be quite noisy, making it a highly challenging problem to interpret the data in a biologically meaningful and systematic manner. Compounding to the stochastic nature of the underlying biological processes under investigation, many non-biological technical factors could contribute to the readings of the biological data collected, which could significantly affect our data interpretation. For example, in the case of microarray chips, the readings of microarray signal intensity could be highly sensitive to small differences in local surface properties of a micrarray chip, making it even difficult to reproduce the same data under the same controllable experimental conditions. Our early focus will be to improve experimental designs for microarray experiments for more reliable data collection.

In this project, we will develop improved experimental processes in order to reduce microarray expression variability. We propose to perform a variety of statistically designed experiments to elucidate the error sources for microarray experiments. Yeast microarrays will be used in these initial experiments since most experience with microarrays has resulted from experiments with yeast microarrays. The results from the final optimized microarray experiments will generate information about the error structure of the microarray data. This information will be used to evaluate bioinformatics algorithms by providing a realistic er-

ror structure. In addition, this information will facilitate the use of improved algorithms that require knowledge of the covariance structure of the noise. This will result in improved data quality and reduced dependence on data normalization and preprocessing. As a result, it will be possible to place more confidence in the assumption that the observed variations in the data could be directly attributed to actual biological variation in the sample rather than on the experimental variability as has often been the case with current microarray experiments. Similar studies will be extended to other high-throughput biological data production experiments, to improve the overall quality of our initial data for information extraction.

Improved data mining capabilities will then be developed for various biological data sources. Our initial focus will still be on microarray chip data. Data clustering is a powerful technique for grouping unstructured data sharing common or similar features. It facilitates recognition of patterns shared by some subsets of the data, and identification of significant hidden signals. Data clustering is often used as the first step in mining a large quantity of data. The basic idea of clustering can be stated as follows. For a given set of data points (e.g., multi-dimensional vectors), group the data into clusters so that (1) data of the same cluster have similar (specified) features, and (2) data of different clusters have dissimilar features. We will develop significantly improved capabilities for data clustering, based on the concept of minimum spanning trees (Prim, 1957). Our preliminary studies have established rigorous relationship between data clustering and searching for substrings in a string with certain properties (Xu et al., 2002; Olman et al., 2002). Based on this fundamental relationship, we will develop a suite of data clustering tools capable of solving data clustering problems that are not easily solvable using any existing clustering tools, including problems of identifying/extracting data clusters from a very noisy data set, etc. Such improved data mining capabilities should allow us to more effectively extract significant patterns and co-relations from a large quantity of unstructured data, to be used for pathway inference.

Parallel efforts will be given to developments of more effective capabilities for mining genomic sequence data for inference of operon structures, identification of orthologs, and identification of protein binding sites, for the purpose of deriving possible parts list (which genes may be involved in a pathway) of a target pathway. The "parts" information, along with the possible protein-protein interaction information, causality information in a gene network, and identified co-regulated gene list, will provide a set of basic components for computational capabilities for pathway inference.

Improved computational capabilities for inference of biological pathways

Our proposed pathway inference framework will consist of the following main components:

1. Prediction of potential genes/proteins involved in a specific pathway
2. Function assignment of a protein
3. Identification of co-regulated genes and interacting proteins
4. Mapping proteins to a known biological pathway
5. Inference of a specific pathway consistent with available information

Prediction of potential genes/proteins involved in a specific pathway. The first step towards pathway construction is to select a relatively small set (<100) of genes/proteins (*list A*) that are potentially involved in the target pathway. To do this, we will first use gene expression profiles to find highly induced genes under the condition that activates the pathway. For example, in studying leucine transport pathway in yeast, we can monitor the up-regulated genes under leucine rich environment. Some of the induced genes are probably involved directly in the pathway; others may be indirectly involved (e.g., general energy pathway) or not relevant at all (due to noise in experimental data). Then we will use protein-protein interaction/co-regulated gene information (see the following) to add more genes to this list. Any proteins that interact directly with a protein in list A will be selected. In addition, proteins that are connected through a series of protein-protein interactions in list A will also be selected. This will form the list of the potential components for the target pathway. We will also use homologous information to build this list. If the homologs of two proteins in another organism interact with each other, the two proteins may interact with each other in *Synechococcus* as well. A user can review this list and add/remove the genes based on his/her own knowledge.

Function assignment of a protein. Some genes in the selected list may not have function assignment. We will use gene expression data, protein-protein interaction data, and co-regulation/regulon data to infer a possible functional role of a hypothetical protein. The idea is "guilt-by-association": if a hypothetical protein interacts with two or more proteins with the same function, it is likely that the hypothetical protein also has this function. If we have determined a hypothetical protein in a regulon and other genes in the regulon have the same known cellular role, then this hypothetical protein may have the same cellular role as well. In addition, we can study genes with similar expression profiles to the hypothetical protein. If there is a consensus among the functions of these genes, it may suggest the function of the hypothetical protein.

Identification of co-regulated genes and interacting proteins. To determine co-regulated genes, we will first cluster gene expression profiles using the selected list. Then we will retrieve upstream regions of these genes. If the co-expressed genes share similar sequence-motifs in their upstream regions, we will assign them as co-regulated genes. Also genes predicted to be in the same operon will also be added to the list of co-regulated genes. After identifying co-regulated genes, we will determine which protein–protein interactions determined above are true interactions (it is expected some of them are false interactions). If two proteins are co-regulated, their interaction determined above is probably true. If a protein–protein interaction is predicted from multiple sources, this interaction pair should have higher likelihood to be true.

Mapping proteins to a known biological pathway. We will search the selected list of proteins against all known biological pathways. This effort will benefit from the literature mining tools also developed in this subproject. If a significant fraction of the proteins can map to a known pathway of a different organism (through homology), a pathway can be constructed for these proteins using the known pathway as a template. The unmapped parts of the pathway may suggest what other proteins we should look in *Synechococcus* genome.

For each of the identified component protein or protein-protein interaction, using our own or others' tools, we will first thoroughly evaluate the reliability of the prediction, based on well-characterized gene/protein functions, protein–protein interactions, operons/regulons. These reliability values will be used in our inference framework when putting all pieces together.

Inference of a specific pathway consistent with available information. Our initial objective is to work on regulatory pathways that may have similar or related known pathways in other genomes. This way, we will have at least a partial template to work with. As our capabilities mature, we will gradually move to more *ab initio* constructions of pathways. In the template-based case, the key issue is to identify all the corresponding parts (components and interactions) in the template pathway and the missing parts in *Synechococcus*. Then the system will ask for either new data to be generated by further experiment or/and for a revised template pathway by the user. In the more *ab inito* construction of a pathway, the system will need some general guiding information from the user like what functional classes of proteins may be involved, how large the target pathway is approximately, which known genes/proteins are probably involved, etc. Then inference framework should suggest a list of possible models based on the general information extracted from various data sources and specific information about the target pathway. We will explore a number of possible computational approaches as the main tool for constructing such frameworks, including Bayesian networks (Friedman et al., 2000), expert system techniques (Takai-Igarashi and Kaminuma, 1999), and combinatorial optimization approaches (Kyoda et al., 2000). These methods have been applied to infer regulatory networks using single source of information. They were only successful on artificial simulated data but not on real biological data, since a single source of data is insufficient to derive regulatory networks as discussed above. Here, by combining different sources of information, we expect to make significant advances in reference of regulatory networks.

4.0 SYSTEMS BIOLOGY MODELS FOR *SYNECHOCOCCUS* SP.

Ultimately, all of the data that is generated from experiment must be interpreted in the context of a model system. Individual measurements can be related to a very specific pathway within a cell, but the real goal

is a systems understanding of the cell. Given the complexity and volume of experimental data as well as the physical and chemical models that can be brought to bear on subcellular processes, systems biology or cell models hold the best hope for relating a large and varied number of measurements to explain and predict cellular response. Clearly, cells fit the working scientific definition of a complex system: a system where a number of simple parts combine to form a larger system whose behavior is much harder to understand. The primary goal of this subproject is to integrate the data generated from the overall project's experiments and lower level simulations, along with data from the existing body of literature, into a whole cell model that captures the interactions between all of the individual parts. It is important to note here that all of the information that is obtained from the previously described efforts in this project is vital to the work here. In a sense, this is the "Life" of the "Genomes to Life" theme of this project.

The precise mechanism of carbon sequestration in *Synechococcus* is poorly understood. There is much unknown about the complicated pathway by which inorganic carbon is transferred into the cytoplasm and then converted to organic carbon. While work has been carried out on many of the individual steps of this process, the finer points are lacking, as is an understanding of the relationships between the different steps and processes. Understanding the response of *Synechococcus* to different levels of CO_2 in the atmosphere will require a detailed understanding of how the carbon concentrating mechanisms in *Synechococcus* work together. This will require looking these pathways as a system.

The aims of this section are to develop and apply a set of tools for capturing the behavior of complex systems at different levels of resolution for the carbon fixation behavior of *Synechococcus*. We briefly describe here those 4 objectives.

Protein interaction network inference and analysis using large-scale experimental data and simulation results

Experimentally, regulatory networks are generally probed with microarray experiments, and protein network interactions have been investigated with 2-hybrid screening. All the inference computational techniques have so far been based on probabilistic frameworks that search the space of all possible labeled graphs. Our aim is to infer networks from multiple sources including phage display experimental data and simulation results from subprojects 2 and 3. Furthermore, instead of searching networks in the space of all labeled graphs we proposed to search networks in the space of scale-free graphs. The scale-free nature of protein networks was first discovered by Jeong et al. (Jeong, 2000) and independently verified by Gomez et al. (Gomez, 2001). Since the number of scale-free graphs is many order of magnitudes smaller than the number of labeled graphs we expect to develop a method far more efficient than the current state-of-the art.

It is well established that proteins interact through specific domains. While many proteins are composed of only one domain, multiple domains are also present and must be considered when reconstructing networks (Uetz, 2000). Probabilities of attraction between protein domains have been derived from phage display data (Tong, 2002), and protein–protein interaction databases. Note that probability of attraction between domains can be calculated from binding energies computed through molecular simulations in subproject 2. Considering two multi-domains proteins i and j, one can then define a probability (p_{ij}) of attraction between these proteins as (Gomez, 2002):

$$p_{ij} = \sum_{d_m \in \nu_i} \sum_{d_n \in \nu_j} \frac{p(d_m,d_n)}{|\nu_i||\nu_j|} \tag{1}$$

where v_i (v_j) is the domain set of protein i (j), and $p(d_m,d_n)$ is the probability of attraction between domains d_m and d_n. Thus, the problem of inferring a protein-protein interaction network from domain–domain interaction probabilities reduces to finding a graph $G = (V,E)$ where the vertices of V are proteins and the edges of E are protein-protein interactions that maximizes the probability:

$$P(E) = \prod_{e_{ij} \in E} p_{ij} \prod_{e_{kl} \notin E} (1 - p_{kl}) \tag{2}$$

The trivial solution to this problem, which consists of selecting only the edges with probability of >0.5 is not appropriate because protein–protein interaction networks are generally scale-free networks [11], which is an additional constraint not captured in Eq. 4-2.

Our proposed work is composed of the four following steps:

1. Develop methodology to characterize and analyze scale-free networks and protein interaction networks
2. Compute domain–domain attraction probabilities from phage display data, molecular simulations, and protein–protein interaction databases
3. Sample scale-free networks that maximize $P(E)$ computed in step 2 using labeled graph sampling algorithm and characteristics developed in step 1
4. Compare predicted networks with experimentally derived two-hybrid networks. Adjust domain–domain attraction probabilities and repeat steps 2–4 until agreement between predicted and two-hyrid networks is reached.

The above four tasks will be tested with the yeast proteome for which there is already ample data and then will be applied to *Synechococcus* when experimental and simulation data become available.

Discrete component simulation model of the inorganic carbon to organic carbon process

Once protein networks have been inferred, one can then study their dynamics. While even the simplest prokaryotic cells are extremely complex, this complexity is generally driven by a relatively small number of unique cellular components. One of the consequences of this is that many important processes in cells can be controlled by the interaction of a very small number of individual reactants. This can lead to a wide range of different behaviors associated with cells of identical type due to the fluctuations in the number and position of its reactants. In many cases, it is important to understand how this randomness affects cell behavior through computer modeling.

There are two different ways in which the individual particle method can be implemented. In the first model, "reactions" are calculated by a stochastic method such as that described by Gillespie (Gillespie, 1976) with recent developments by Gibson and Bruck (Gibson, 2000). In this method, there is a set of possible reactions that can occur given the products that exist. There are also reaction rates that are associated with any of these events occurring. For calculations where spatial details are more important, a second model is used that is a little more sophisticated. In this model, each of the objects is modeled separately and its spatial position tracked separately, in the spirit of the `Mcell` code by Stiles and Bartol (Stiles, 2001). (We note here for clarity that the "particles" described in this section are not atoms or even necessarily molecules, but simply individual objects in the cell that must be tracked separately.)

There are two primary tasks associated with this objective:

Stochastic method. We will first build a serial version of the code, based on the work that has already been done by Lok and Brent at tMSI. We will test this code on yeast data, and *Synechococcus* data from other subprojects when it becomes available. In this serial version we will address the event scheduling issues related to the sub-volume partitioning so that the debugging processes will be more straightforward than it would be on the parallel version. After the serial code is working, we will begin to develop a massively parallel version of this code based on domain decomposition.

Individual particle method. This method will begin by adapting Sandia's existing particle code (ICARUS) to work on biological systems. Boundary conditions will be implemented that allow reactions on the interfaces. This will model biological processes that occur on the cell membrane and the surfaces of internal structures. The ultimate goal is to be able to handle more than 10^7 individual particles using hundreds of processors.

In both models we can start with a higher concentration of inorganic carbon near the membrane and then run the model forward in time to generate a simulation of how the inorganic and organic carbon (in the carboxysomes) coexist inside the cell. Once the network is set up, one can then change individual reactant amounts or reaction rates and test to see how this affects the results. Finally, these techniques can be used to help determine unknown variables in the network by comparing the results against experimentally determinable quantities.

Continuous species simulation of ionic concentrations

While a discrete particle simulation is useful for situations where there is a relatively small number of particles, once the concentration of a particular species becomes large enough the discrete method becomes impractical and unnecessary. In this case, the particle number is large enough that the overall behavior is better understood as a continuous phenomenon, where the particle concentration is modeled as a continuous function of space and time. The interactions between various species are described in terms of partial differential equations, and the resulting formulae belong to a general class of equations known as *reaction/diffusion* equations.

One code used to solve the reaction/diffusion equations essential for this effort is a widely used production code at Sandia called MPSalsa (Shadid, 1997). This code has been shown to successfully scale to more than 1,000 processors with very little loss of speed. We plan on using a version of MPSalsa to perform much of the work proposed here.

Despite much research, there is still not a clear consensus on the mechanism by which inorganic carbon is transported across the cell membrane (Kaplan, 1999). There are many mechanisms that are being considered. The simplest is that it passes through the cell membrane as CO_2, and this behavior has been well documented in many microbes. It is also believed that HCO_3^- is actively transported across the membrane via either an ion gradient or by an ATP fueled pump. There is now increasing belief that there may be multiple mechanisms for getting inorganic carbon into the cytoplasm. Some of the CO_2 that exists in the cytoplasm is converted into HCO_3^-. When the HCO_3^- reaches the carboxylation site, it is converted to CO_2, which is then used by RuBisCO to form 3-phosphoglycerate (PGA).

The specific goal in this aim is to study interplay between CO_2 and HCO_3. The first work will be done making minor modifications to the existing code (MPSalsa) that allow for species to be created at interfaces to help model specific biological mechanisms, such as membrane transport. Eric Jakobsson and his co-workers at UIUC have done extensive modeling of membranes and ion channels. They will be providing support to this project by modeling proposed ion channel structures based on sequence data to help formulate the boundary conditions for the for inorganic carbon species formulation. The boundary conditions on the simulation can be set to represent both the steady diffusion of CO_2 across the membrane, and point sources of HCO_3^- related to specific pumps located in the membrane. The carboxylation site could also be modeled as a sink for HCO_3^- and a source for CO_2 and PGA.

Once the work gets done obtaining all of the boundary conditions regarding inorganic carbon transport, the simulation will be used to study what concentrations of carbon could be sequestered given various known kinetic constants associated with RuBisCO (as discussed in section 1.0) and membrane transport. We will then compare our results to experimental measurements obtained in this proposal and elsewhere, and use this to drive the direction of future experiments.

Synechococcus *carboxysomes and carbon sequestration in bio-feedback, hierarchical modeling*

This aim answers the questions, "How does one utilitize genomic and proteomic information in problems manifest at the ecosystem level?" To begin, consider a conceptual organization that can be modeled via a hierarchical, object-oriented design, whereby conceptually discrete systems are linked by levels of interaction. Details of each level are handled within a "black box," communicating to levels above and below by specified rules based on scientifically known or hypothesized mechanisms of interaction. The model allows the connection of levels by the imposition of *de novo* laws as discovered in the respective disciplines, their rules of interaction and axiomatic behavior, as well as an actual examination of the state of the level.

As a demonstrative example of the utility of a hierarchical, feedback model, we have tested the preliminary implementation in the complex scenario of the genetic basis of flu pandemics. Influenza is a negative-stranded RNA virus of the family *Orthomyxoviridae.* Importantly, each pandemic has been associated with the discovery of a new serotype for the virus' hemagglutinin (HA) protein. Swine, and particularly birds, serve as reservoirs for the HA subtypes. With this basic knowledge of genetic factors underlying influenza's virulence, we now seek factors that create HA variation. RNA-RNA recombination is known in numerous viruses, including influenza (for review, see Worobey and Holmes 1999). Using our hierarchical model, we

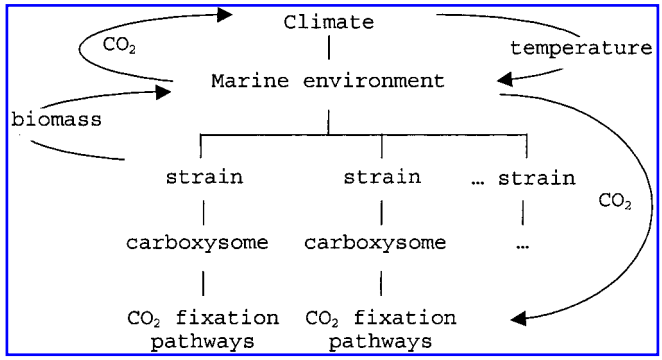

FIG. 1. Hierarchical model relating pathways to carbon cycling.

demonstrated the greatly increased morbidity associated with the RNA-RNA recombination model (as opposed to no RNA-RNA recombination).

To investigate the importance of *Synechococcus* in carbon cycling using a data-driven, hierarchical model, we seek to directly incorporate genomic and proteomic knowledge of *Synechococcus* to understand how conditions, such as a 1°C increase in ambient temperature, affect carbon fixation of important and ubiquitous marine populations (Fig. 1). We propose to do this by underlaying the carboxysome of Figure 2 with known carbon fixation metabolic pathway information such as that available at http://genome.ornl.gov/keggmaps/syn_wh/07sep00/html/map00710.html. The network dynamics of the previous sections of this proposal give us a model of carbon fixation dependent on a variety of external parameterizations, such as ambient water temperature, CO_2 diffusion rates, and *Synechococcus* growth rates.

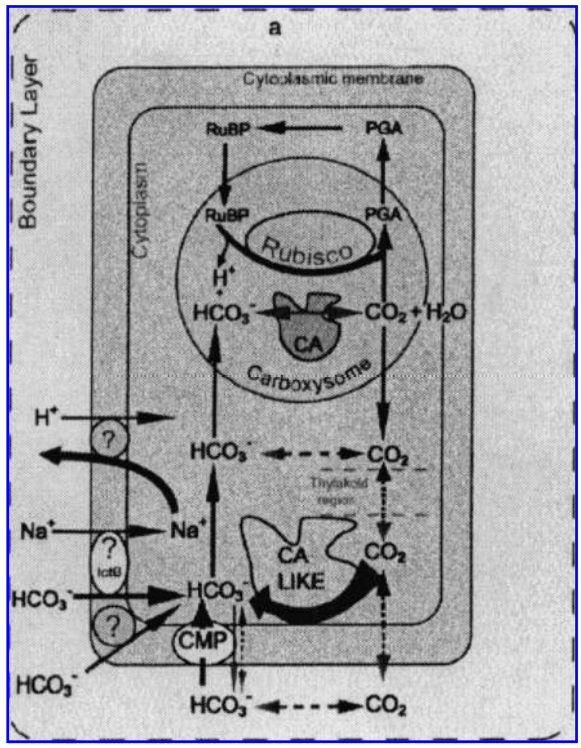

FIG. 2. Carbon concentrating mechanism (from Kaplan and Reinhold, 1999).

A broader result of this work on *Synechococcus* is to help us understand how biological reactions to environmental conditions feedback onto the environmental conditions themselves: thus the loop back in Figure 1 between CO_2 affecting growth rates and marine biomass, which in turn affect carbon sequestration. The strains in Figure 1 each encapsulate a variant in CO_2 fixation pathways as similarly used in the previous worked example.

5.0 COMPUTATIONAL BIOLOGY WORK ENVIRONMENTS AND INFRASTRUCTURE

The explosion of data being produce by high-throughput experiments will require data analysis tools and models which are more computationally complex, more heterogeneous, and require coupling to enormous amounts of experimentally obtained data in archived ever changing formats. Such problems are unprecedented in high performance scientific computing and will easily exceed the capabilities of the next generation (PetaFlop) supercomputers.

The principal finding of a recent DOE Genomes to Life (GTL) workshop was that only through computational infrastructure dedicated to the needs of biologists coupled with new enabling technologies and applications will it be possible "to move up the biological complexity ladder" and tackle the next generation of challenges. This section discusses the development of a number of such capabilities including work environments such as electronic notebooks and workflow environments, and high performance computational systems to support the data and modeling needs of GTL researchers, particularly those involved in this *Synechococcus* study.

Important to enabling technologies is the issue of ease of use and coupling between geographically and organizationally distributed people, data, software, and hardware. Today most analysis and modeling is done on desktop systems, but it is also true that most of these are greatly simplified problems compared to the needs of GTL. Thus an important consideration in the GTL computing infrastructure is how to link the GTL researchers and their desktop systems to the high performance computers and diverse databases in a seamless and transparent way. We believe that this link can be accomplished through work environments that have simple web or desktop based user interfaces on the front-end and tie to large supercomputers and data analysis engines on the back-end.

These work environments have to be more than simple store and query tools. They have be conceptually integrated "knowledge enabling" environments that couple vast amounts of distributed data, advanced informatics methods, experiments, and modeling and simulation. Work environment tools such as the electronic notebooks have already shown their utility in providing timely access to experimental data, discovery resources and interactive teamwork, but much needs to be done to develop integrated methods that allow the researcher to discover relationships and ultimately knowledge of the workings of microbes. With large, complex biological databases and a diversity of data types, the methods for accessing, transforming, modeling, and evaluating these massive datasets will be critical. Research groups must interact with these data sources in many ways. In this effort, we will develop a problem solving environment with tools to support the management, analysis, and display of these datasets. We will also develop new software technologies including "Mathematica"-type toolkits for molecular, cellular and systems biology with highly optimized life science library modules embedded into script-driven environments for rapid prototyping. These modules will easily interface with database systems, high-end simulations, and collaborative workflow tools for collaboration and teaching.

Working environments: the lab benches of the future

This project will result in the development of new methods and software tools to help both experimental and computational efforts characterize protein complexes and regulatory networks in *Synechococcus*. The integration of such computational tools will be essential to enable a systems-level understanding of the carbon fixation behavior of *Synechococcus*. Computational working environments

will be an essential part of our strategy to achieve the necessary level of integration of such computational methods and tools.

Because there is such diversity among computational life science applications in the amount and type of their computational requirements, the user interface designed in this effort will be designed to support three motifs. The first is a biology web portal. These have become popular over the past three years because of their easy access and transparent use of high performance computing. One such popular web portal is ORNL's *Genome Channel.* The Genome Channel is a high-throughput distributed computational environment providing the genome community with various services, tools, and infrastructure for high quality analysis and annotation of large-scale genome sequence data. We plan to leverage this existing framework, which is based on the Genomes Integrated Supercomputer Toolkit (GIST), to create a web portal for the applications developed in this proposal.

The second motif is an electronic notebook. This electronic equivalent of the paper lab notebook is in use by thousands of researchers across the nation. Biology and Pharma labs have shown the most interest in this collaboration and data management tool. Because of its familiar interface and ease of use, this motif provides a way to expose reluctant biologists to the use of software tools as a way to improve their research. The most popular of the electronic notebooks is the ORNL enote software. This package provides a very generic interface that we propose to make much more biology centric by integrating the advanced bioinformatics methods described in this proposal into the interface. In out years we plan to incorporate metadata management into the electronic notebook to allow for tracking of data pedigree, etc.

The third motif will be a *Matlab*-like toolkit whose purpose would be fast prototyping of new computational biology ideas and allow for a fast transition of algorithms from papers into tools that can be made available to an average person sitting in the lab. No such tool exists today for biology.

For all three of the working environment motifs we will build an underlying infrastructure to: (1) support new core data types that are natural to life science, (2) allow for new operations on those data types, (3) support much richer features, and (4) provide reasonable performance on typical life science data. The types of data supported by electronic notebooks and problem solving environments should go beyond sequences and strings and include trees and clusters, networks and pathways, time series and sets, 3D models of molecules or other objects, shapes generator functions, deep images, etc. Research is needed to allow for storing, indexing, querying, retrieving, comparing, and transforming those new data types. For example, such tools should be able to index metabolic pathways and apply a comparison operator to retrieve all metabolic pathways that are similar to a queried metabolic pathway.

Creating new GTL-specific functionality for the work environments

There are a large number of GTL-specific functionalities that could be added to the work environments. For this effort we have selected three that have wide applicability across computational biology to illustrate how the work environments can be extended. These are as follows:

1. Graph data management for biological network data
2. Efficient data organization and processing of microarray databases
3. High performance clustering methods

Graph data management for biological network data. In this first area, we will develop general-purpose graph-based data management capabilities for biological network data produced by this *Synechococcus* effort as well as from other similar efforts. Our system will include an expressive query language capable of encoding select-project queries, graph template queries, regular expressions over paths in the network, as well as subgraph homomorphism queries (e.g. find all of examples of pathway templates in which the enzyme specification is a class of enzymes). Such subgraph homomorphism queries arise whenever the constraints on the nodes of the query template are framed in terms of generic classes (abstract noun phrases) from a concept lattice (such as the Gene Ontology), whereas the graph database contents refer to specific enzymes, reactants, etc. Graph homomorphism queries are known to be *NP*-hard and require specialize tech-

niques that cannot be supported by translating them into queries supported by conventional database management systems.

This work on graph databases is based on the premise that such biological network data can be effectively modeled in terms of labeled directed graphs. This observation is neither novel nor controversial: a number of other investigators have made similar observations (e.g. the AMAZE database, VNM00). Other investigators have suggested the use of stochastic Petri Nets (generally described by Directed Labeled Graphs) to model signaling networks. Some nodes represent biochemical entities (reactants, proteins, enzymes, etc.) or processes (e.g. chemical reactions, catalysis, inhibition, promotion, gene expression, input-to-reaction, output-from-reaction, etc.). Directed edges connect chemical entities and biochemical processes to other biochemical processes or chemical entities. Undirected edges can be used to indicate protein interactions.

Current systems for managing such network data offer limited query facilities, or resort to ad hoc procedural programs to answer more complex or unconventional queries, which the underlying (usually relational) DBMSs can not answer. The absence of general purpose query languages for such graph databases either constrains the sorts of queries biologists may ask, or forces them to engage in tedious programming whenever they need to answer such queries. For these reasons, we will focus our efforts on the development of the graph query language and a main memory query processor. We plan to use a conventional relational DBMS for the persistent storage (perhaps DB2, which supports some recursive query processing). The main memory graph query processor will directly call the relational database management system (*i.e.,* both will reside on the server). The query results will be encoded (serialized) into XML and a SOAP-based query API will be provided, to permit applications or user interfaces to run remotely.

We will also explore the use of graph grammars for describing query languages, network data, and the evolution of biological networks. Graph grammars are the graph analog of conventional string grammars. Thus the left hand side of a GG rule is generally a small graph, whereas the right hand side of the GG rule would be a larger (sub-) graph. Graph grammars can be used for graph generation (e.g. to model network evolution) and graph parsing. They have been used to describe various sorts of graph languages. Graph grammars could be useful for specifying the graph query language powerful enough for the graph operations described above.

Efficient data organization and processing of microarray databases. Microarray experiments have proven very useful to functional genomics and the data generated by such experiments is growing at a rapid rate. While initial experiments were constrained by the cost of microarrays, the speed with which they could be constructed, and occasionally by the sample generation rates, many of these constraints have been or are being overcome. One could readily envision that data production rates might increase another factor of 5 or 10. Note that we are concerned here with the processed data, not the much larger raw image data, which we assume will likely not be kept in a DBMS. Datasets of 50 or 100 GB/year $\times$ 3 or 4 years exceed likely main memory configurations. This does not even account for record overhead, or indices. It is likely that most of this data will be kept on disk. Thus we will need efficient database designs, indices and query processing algorithms to retrieve and process these large datasets from disk.

It will be necessary to support queries over these relations in combination with the spot data in order to permit queries that are meaningful to the biologists. Note that the space described above represents the Cartesian product of the experimental conditions and the genes. However, we can expect replication of spots and experiments, since replication is essential to reliable statistical analysis of this very noisy data. In addition, it will be necessary to support ontology of the various genes, gene products and a biological network database describing various cellular processes (metabolic, signal transduction, gene regulation).

A common query might ask (over some subset of the experimental design) which genes are overexpressed relative to their expression for standard experimental conditions. Other queries might request restricting the set of genes considered to certain pathways or retrieving pathways in addition to genes. To support such queries, it is necessary to join the results of conditions on the experimental design with the microarray spot data in order to identify the genes that are overexpressed. This implies the capability of searching over one or more of the spot attributes.

Indexing over a billion or more elements is a daunting task. Conventional indexing techniques provided by commercial database systems, such as B-trees, do not scale. One of the reasons for this is that general-

purpose indexing techniques are designed for data that can be updated over time. Recognizing this problem, other indexing techniques have been proposed, notably techniques that take advantage of the static nature of the data, as is the case with much of scientific data resulting from experiments or simulations. One of the most effective methods of dealing with large static data is called "bitmap indexing" (Wu et al., 2001; Wu et al., 2002). The main idea for bitmap indexing is to partition each attribute into some number of bins (such as 100 bins over the range of data values), and to construct bitmaps for each bin. Then one can compressed the bitmaps and perform logical operations to achieve a great degree of efficiency.

LBNL has developed highly efficient bitmap indexing techniques that were shown to perform one to two orders of magnitude better than commercial software, and where the size of the indexes are only 20–30% the size of the original vertical partition (Wu et al., 2001; Wu et al., 2002). To achieve this we have developed specialized compression techniques and encoding methods that permit the logical operation to be performed directly on the compressed data. We have deployed this technique in a couple of scientific applications, where the number of elements per attribute vector reaches hundreds of millions to a billion elements. We propose here to use this base of software to the problem of indexing microarray spot data.

High performance clustering methods. In this third area, we will be implementing a clustering algorithm named RACHET into our work environments. RACHET builds a global hierarchy by merging clustering hierarchies generated locally at each of the distributed data sites and is especially suitable for very large, high-dimensional, and horizontally distributed datasets. Its time, space, and transmission costs are at most linear in the size of the dataset. (This includes only the complexity of the transmission and agglomeration phases and does not include the complexity of generating local clustering hierarchies.)

Clustering of multidimensional data is a critical step in many fields including data mining, statistical data analysis, pattern recognition and image processing. Current popular clustering approaches do not offer a solution to the distributed hierarchical clustering problem that meets all these requirements. Most clustering (Murtagh, 1983; Day and Edelsbrunner, 1984; Jain et al., 1999) are restricted to the centralized data situation that requires bringing all the data together in a single, centralized warehouse. For large datasets, the transmission cost becomes prohibitive. If centralized, clustering massive centralized data is not feasible in practice using existing algorithms and hardware. RACHET makes the scalability problem more tractable. This is achieved by generating local clustering hierarchies on smaller data subsets and using condensed cluster summaries for the consecutive agglomeration of these hierarchies while maintaining the clustering quality. Moreover, RACHET has significantly lower (linear) communication costs than traditional centralized approaches.

In summary, this project will require an infrastructure that enables easy integration of new methods and ideas and supports biology collaborators at multiple sites so they can interact as well as access to data, high performance computation, and storage resources.

ACKNOWLEDGMENTS

We are grateful to the vision and support of the DOE Office of Science, the sponsor of the Genomes to Life program. The DOE Genomes to Life (GTL) program is unique in that it calls for "well-integrated, multidisciplinary (e.g. biology, computer science, mathematics, engineering, informatics, biophysics, biochemistry) research teams," with strong encouragement to "include, where appropriate, partners from more than one national laboratory and from universities, private research institutions, and companies." Such guidance is essential to the success of the GTL program in meeting its four ambitious goals. To this end, our effort includes participants from four DOE laboratories (Sandia National Laboratories, Oak Ridge National Laboratory, Lawrence Berkley National Laboratory, and Los Alamos National Laboratory), four universities (California/San Diego, Michigan, California/Santa Barbara, and Illinois/Urbana/Champaign), and three institutes (The National Center for Genomic Resources, The Molecular Science Institute, and the Joint Institute for Computational Science).

REFERENCES

BATEMAN, A., BIRNEY, E., CERRUTI, L., et al. (2002). The Pfam protein families database. Nucleic Acids Res. **30,** 276–280.

BEDU, S., LAURENT, B., and JOSET, F. (ed.) (1992). *Membranes and Soluble Carbonic Anhydrase in a Cyanobacterium. Synechocystis PCC6803, Vol. III.* (Kluwer Academic Publishers, Dordrecht, The Netherlands).

CANNON, G.C., and SHIVELY, J.M. (1983). Characterization of a homogeneous preparation of carboxysome from *Thiobacillus neapolitanus.* Arch. Microbiol. **134,** 52–59.

CANNON, G.C., BRADBURNE, C.E., ALDRICH, H.C., et al. (2001). Microcomartments in prokaryotes: carboxysomes and related polyhedra. Appl. Env. Microbiol. **67,** 5351–5361.

CHANG, G., and ROTH (C.B.) 2001, Structure of MSBA from *E. coli:* a homolog of the multidrug resistance ATP binding cassette (ABC) transporters. Science **293,** 1793–1800.

DANDEKAR, T., SNEL, B., HUYNEN, M., et al. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. **23,** 324–328.

DAY, W., and EDELSBRUNNER, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. J. Classificiation **1,** 7–24.

ENGLISH, R.S., LORBACH, S.C., QIN, X., et al. (1994). Isolation and characterization of a carboxysome shell gene from *Thiobacillus neapolitanus.* Mol. Microbiol. **12,** 647–654.

ENRIGHT, A.J., ILIOPOULOS, I., KYRPIDES, N.C., et al. (1999). Protein interaction maps for complete genomes based on gene fusion events. Nature **402,** 86–90.

FALZONE, C.J., KAO, Y.H., ZHAO, J.D., et al. (1994). Three-dimensional solution structure of PSAE form the cyanobacterium *Synechococcus* sp. strain PCC-7002, a photosystem-i protein that shows structural homology with SH3 domains. Biochemistry **33,** 6052–6062.

FIELDS, S., and SONG, O. (1989). A novel genetic system to detect protein–protein interactions. Nature **340,** 245–246.

FINLEY, R.L., and BRENT, R. (1994). Interaction mating reveals binary and ternary connections between drosophila cell cycle regulators. Proc. Natl. Acad. Sci. USA **91,** 12980–12984.

FRIDLYAND, L., KAPLAN, A., and REINHOLD, L. (1996). Quantitative evaluation of the role of a putative CO_2-scavenging entity in the cyanobacterial CO_2-concentrating mechanism. Biosystems **37,** 229–238.

FRIEDMAN, N., LINIAL, M., NACHMAN, I., et al. (2000). Using Bayesian networks to analyze expression data. J. Comput. Biol. **7,** 601–620.

GARCIA, A.E., and SANBONMATSU, K.Y. (2001). Exploring the energy landscape of a beta hairpin in explicit solvent. Proteins–Structure Funct. Genet. **42,** 345–354.

GAVIN, A.C., BOSCHE M., KRAUSE R., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature **415,** 141–147.

GIBSON, M.A., and BRUCK, J. (2000). Efficient exact stochastic simulation of chemical systems with many species and many channels. J. Phys. Chem. **104,** 1876–1889.

GILLESPIE, D. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J. Comp. Phys. **22,** 403–434.

GOMEZ, S.M., LO S.-H., and RZHETSKY A. (2001). Probabilistic prediction of unknown metabolic and signal-transduction network. Genetics **159,** 1291–1298.

GOMEZ, S.M., and RZHETSKY, A. (2002). Towards the prediction of complete protein–protein interaction networks. R.B. Altman, A.K. Dunker, L. Hunter, K. Lauderdale, and T.E. Klein, eds. *Pacific Symposium on Biocomputing* (Lihue, Hawaii, World Scientific), 413–424.

HU, Y., GRUHLER, A., HEILBUT, A., et al. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature **415,** 180–183.

HOCH, J.A., and SILHAVY T.J. (1995). *Two-component signal transduction.* (Washington, DC, ASM Press).

HOLTHUIJZEN, Y.A., VANBREEMEN, J.F.L., KUENEN, J.G., et al. (1986). Protein composition of the carboxysomes of *Thiobacillus neapolitanus.* Arch. Microbiol. **144,** 398–404.

HUYNEN, M., SNEL, B., LATHE, W., et al. (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res. **10,** 1204–1210.

JAIN, A.K., MURTY, M.N., and FLYNN, P.J., (1999). Data clustering: a review. ACM Computing Surveys **31,** 264–323.

JEONG H., TOMBOR B., ALBERT, R., et al. (2000). The large-scale organization of metabolic networks. Science **407,** 651–654.

KAPLAN, A., FRIEDBERG, R., SCHWARZ, R., et al. (1989). The "CO_2 concentrating mechanism" of cyanobacteria: physiological, molecular and theoretical studies. Photosynth. Res. **17,** 243–255.

KAPLAN, A., and REINHOLD, L. (1999). CO_2 concentrating mechanisms in photosynthetic microorganisms. Annu. Rev. Plant Physiol. Plant Mol. Biol. **50,** 539–570.

KYODA, K.M., MOROHASHI, M., ONAMI, S., et al. (2000). A gene network inference method from continuous-value gene expression data of wild-type and mutants. Genome Inform. **11,** 196–204.

LANARAS, T., HAWTHORNTHWAITE, A.M., and CODD, A. (1985). Localization of carbonic anhydrase in the cyanobacterium *Chlorogloeopsis fritschii.* FEMS Microbiol. Lett. **26,** 285–288.

MARCOTTE, E.M., PELLEGRINI, M., THOMPSON, M.J., et al. (1999). A combined algorithm for genome-wide prediction of protein function. Nature **402,** 83–86.

MARTIN, M.G. (2000). Available: www.cs.sandia.gov/projects/towhee.

MARTIN, M.G., and SIEPMANN, J.I. (1999). Novel configurational-bias Monte Carlo method for branched molecule—transferable potentials for phase equilibria—2. United-atom description of branched alkanes. J. Phys. Chem. B **103,** 4508–4517.

MITSUTAKE, A., SUGITA, Y., and OKAMOTO, Y. (2001). Generalized-ensemble algorithms for molecular simulations of biopolymers. Biopolymers Pept. Sci. **60,** 96–123.

MURTAGH, F. (1983). A survey of recent advances in hierarchical clustering algorithms. Compu. J. **26,** 354–359.

NINFA, A.J., and ATKINSON, M.R. (1995). Control of nitrogen assimilation by the NRI-NRII two component system of enteric bacteria. *Two-Component Signal Transduction.* J.A. Hoch and T.J. Silhavy, eds. (ASM Press, Washington, D.C.).

OLMAN, V., XU, D., and XU, Y. (2002). A new framework for biological data clustering using minimum spanning trees. *Proceedings of the 7th Pacific Symposium on Biocomputing* (in press).

OVERBEEK R., LARSEN N., PUSCH G.D., et al. (2000). WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res. **28,** 123–125.

PELLEGRINI, M., MARCOTTE, E.M., THOMPSON, M.J., et al. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA **96,** 4285–4288.

PRATT, L.A., and SILHAVY, T.J. (1995). Porin regulon in *Esherichia coli. Two-Component Signal Transduction.* J.A. Hoch and T.J. Silhavy, eds. (ASM Press, Washington, D.C.).

PRICE, G.D., COLEMAN, J.R., and BADGER, M.R. (1992). Association of carbonic anhydrase activity with carboxysomes isolated from the cyanobacterium *Synechococcus* PCC7942. Plant Physiol. **100,** 784–793.

PRIM, R.C. (1957). Shortest connection networks and some generalizations. Bell System Tech. J. **36,** 1389–1401.

PUIG, O., CASPARY, F., RIGAUT, G., et al. (2001). The tandem affinity purification method: a general procedure of protein complex purification. Methods **24,** 218–229.

RODI, D.J., JANES, R.W., SANGANEE, H.J., et al. (1999). Screening of a library of phage-displayed peptides identifies human Bcl-2 as a taxol-binding protein. J. Mol. Biol. **285,** 197–203.

SANBONMATSU, K.Y., and GARCIA, A.E. (2002). Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. Proteins–Structure Funct. Genet. **46,** 225–234.

SHADID, J., HUTCHINSON, S., HENNIGAN, G., et al. (1997). Efficient parallel computation of unstructured finite element reacting flow solutions. Parallel Comput. **23,** 1307–1325.

SHIVELY, J.M., LORBACH, S.C., JIN, S., et al. (1996). Carboxysomes: the genes of *Thiobacillus neapolitanus.* Lidstrom, M.E., and Tabita, F.R., eds. *Microbial Growth on C1 Compounds.* (Kluwer, Dordrecht, The Netherlands), 56–63.

SMITH, G.P., and PETRENKO, V.A. (1997). Phage display. Chem. Rev. 97, 391–410.

STEPHANOPOULOS, G., ARISTIDOU, A.A., and NIELSEN, J. 1998. Metabolic engineering. Biotechnol. Bioeng. **58,** 119–120.

STILES, J., and BARTOL, T. (2001). Monte Carlo methods for simulating realistic synaptic microphysiology using Mcell. De Schutter, E., eds. *Computational Neuroscience: Realistic Modeling for Experimentalists* (CRC Press, Boca Raton), 87–127.

TAKAI-IGARASHI, T., and KAMINUMA, T. (1999). A pathway finding system for the cell signaling networks database. In Silico Biol. **1,** 129–146.

TONG, A.H.Y., DREES, B., NARDELLI, G., et al. (2002). A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science **295,** 321–324.

UETZ, P., GIOT, L., CAGNEY, G., et al. (2000). A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae.* Nature **403,** 623–627.

VAN HELDEN, J., NAIM, A., MANCUSO, R., et al. (2000). Representing and analysing molecular and cellular function using the computer. Biol. Chem. **381,** 921–935.

WOROBEY, M., and HOLMES, E.C. (1999). Evolutionary aspects of recombination in RNA viruses. J. Gen. Virol. **80,** 2535–2543.

WU, K., OTOO, E.J., and SHOSHANI, A. (2001). A performance comparision of bitmap indexes. ACM Int. Conf. Information Knowledge Management 559–561.

WU, K., OTOO, E.J., and SHOSHANI, A. (2002). *Compressed Bitmap Indexes for Faster Search Operations.* LBNL Technical Report, LBNL-49627.

XU, Y., OLMAN, V., and XU, D. (2002). A graph-theoretical approach for gene expression data analysis: an application of minimum spanning trees. Bioinformatics **18,** 526–535.

Address reprint requests to:
Dr. Grant S. Heffelfinger
Sandia National Laboratories
Bldg 701/2101, MS-0885
1515 Eubank SE
Albuquerque, NM 87123

E-mail: gsheffe@sandia.gov

This article has been cited by:

1. Shouqiang Cheng, Yu Liu, Christopher S. Crowley, Todd O. Yeates, Thomas A. Bobik. 2008. Bacterial microcompartments: their properties and paradoxes. *BioEssays* **30**:11-12, 1084-1095. [CrossRef]
2. Dylan M. Morris, Grant J. Jensen. 2008. Toward a Biomechanical Understanding of Whole Bacterial Cells. *Annual Review of Biochemistry* **77**:1, 583-613. [CrossRef]
3. G.X. Yu, E.M. Glass, N.T. Karonis, N. Maltsev. 2006. Knowledge-based voting algorithm for automated protein functional annotation. *Proteins: Structure, Function, and Bioinformatics* **61**:4, 907-917. [CrossRef]