Sandia National Laboratories

13

Carbon Sequestration in *Synechococcus*: A Computational Biology Approach to Relate the Genome to Ecosystem Response

Grant S. Heffelfinger* (gsheffe@sandia.gov)

Sandia National Laboratories, Albuquerque, NM

This talk will provide an update on the progress to date of the Genomics:GTL project led by Sandia National Laboratories: "Carbon Sequestration in *Synechococcus* sp.: From Molecular Machines to Hierarchical Modeling." This effort is focused on developing, prototyping, and applying new computational tools and methods to elucidate 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. While much of our recent progress and results will be presented in detail in the seven or more posters submitted to this meeting (see Davidson et al., Geist et al., Martino et al., Plimpton et al., Samatova et al. Sinclair et al., Xu et al. and others), this talk will recap the larger focus and recent results of the project. Our project's results include experimental data on the *Synechococcus* carboxysome and CO_2 levels on growth rate and protein expression patterns in *Synechococcus* (sp. WH8102), the first characterizations of components of the proteome, and characterizations of the phosphorus and nitrogen regulatory pathways in conjunction with computationally derived predictions of these pathways. Our computational tool development efforts relative to processing high throughput experimental data have yielded new methods and algorithms for gene expression array analysis and a radically new tandem MS, MS/MS data analysis method which enables prototype assignment for large and diverse data sets (~60,000 spectra) with a surprising level of confidence. We have also developed and prototyped new computational tools for microbial systems biology, including methods for multi-scale characterization of protein interactions, methods for recognizing protein functional sites, and an integrating framework such tools. In addition to our work with *Synechococcus* Sp., we have applied these tools to other microbes in collaboration with other Genomics:GTL projects including the ORNL-PNNL microbial proteomics effort for *Rhodopseudomonas palustris* (with F. Larimer and H. McDonald). Our efforts to develop and apply modeling and simulation tools have yielded structural insight into the specificity of the carbon fixing enzyme RuBisCO as well as a computational capability to track spatial and temporal variations in protein species concentrations in realistic cellular geometries for important cyanobacterial subcellular processes. Finally, we have constructed an integrated data infrastructure which allows advanced search and queries across a large, diverse set of data sources (e.g. databases of sequence, structure, pathway, protein interaction, and raw mass spectra and microarray data). Our "*Synechococcus* Encyclopedia," contains all currently available database knowledge about this microbe and we are working now to create an encyclopedia for *Rhodopseudomonas palustris* and *Shewanella* for use by the GTL Microbial Complexes Pipeline project and *Shewanella* Federation respectively. More detailed discussions of our results may be found in our project's quarterly reports, available at www.genomes-to-life.org.

* Presenting author

14

Integrating Heterogeneous Databases and Tools for High Throughput Microbial Analysis

Nagiza Samatva* (samatovan@ornl.gov), Al Geist, Praveen Chandramohan, and Ramya Krishnamurthy

Oak Ridge National Laboratory, Oak Ridge, TN

Going beyond simple data archiving and retrieval of diverse data sets, we will describe a knowledge infrastructure that provides capabilities far beyond what has been available before. As part of the Genomics:GTL *Synechococcus*: From Molecular Machines to Hierarchical Modeling project, we have developed the technology needed to construct an integrated data infrastructure that allows advanced search and queries across a large, diverse set of data sources including sequence databases (COG, INTERPRO, SWISS PROT, TIGR, JGI, PFAM, PRODOM, SMART), structure databases (PDB, COILS, SOSUI, PROSPECT), pathway databases (KEGG), protein interaction databases (BIND, DIP, MIPS), and databases of raw mass spec and microarray data. Both a query language and integrated schema technology were developed to allow search and queries across these diverse databases.

We used our integrated data infrastructure to create a *Synechococcus* Encyclopedia (see Figure 1) containing all the database knowledge in the world about this microbe. This knowledgebase involves the integration of 23 different databases and is being used to do protein complex predictions, and pathway predictions. The technology can be used to create knowledgebases for other organisms and we have begun discussions with other GTL projects about setting up encyclopedias for their microbes.

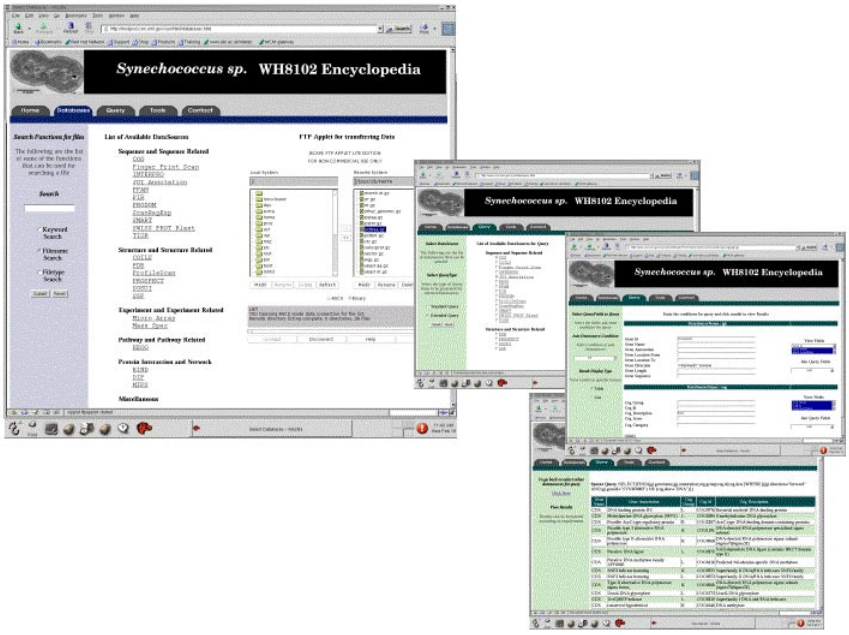
The encyclopedia not only has data but also tools to analyze the data. This past year we have added a suite of easy-to-use web-based analysis tools to the encyclopedia. These tools, which are being developed within our GTL project, include protein function characterization, protein structure prediction, comparative analysis of protein-protein interfaces, metadata entry and browsing, pathway prediction, and electronic notebooks. Several of these tools provide transparent access to supercomputers at ORNL and around the nation. We will describe how the encyclopedia data and analysis tools were used to correctly predict the proteins making up a known membrane complex – including the membrane proteins involved – a task that is presently impossible by experimental mass spec analysis alone.

The new ability to rapidly construct advanced queries that require correlating and combining data from sequence annotations, protein structure, and interaction databases and to use the results in co-located analysis tools allows biologists to combine knowledge and see relationships that were previously obscured by the distributed nature and diverse data types in the biological databases.

The presentation will include "live" demonstrations of advanced queries of the *Synechococcus* Encyclopedia.

Acknowledgement: This project is supported by the U.S.Department of Energy's Genomics:GTL Program under project "Carbon Sequestration in *Synechococcus sp*: From Molecular Machines to Hierarchical Modeling" (http://www.genomes-to-life.org)

Figure 1. *Synechococcus sp*. Encyclopedia. Advanced query and analysis interface. Search all *Synechococcus* databases. Browse experimental and analysis data. Download datasets.



15

Toward Comprehensive Analysis of MS/MS Data Flows

Andrey Gorin* (agor@ornl.gov), Nikita D. Arnold, Robert M. Day, and Tema Fridman

Oak Ridge National Laboratory, Oak Ridge, TN

Tandem mass spectrometry (MS/MS) is a powerful tool applied across several Genomics:GTL projects for a variety of challenging proteomics projects: search for modified proteins, characterization of whole cell proteome, and identification of components of protein molecular machines. Despite great variety of the biological drivers, computational algorithms used "under the hood" face exactly the same challenges, and existing limitations of such algorithms are reproduced across many experimental designs. in ion trap devices under common conditions only ~20% of MS/MS spectra lead to peptide identifications that are worth to be considered, and misidentification rates remains to be high.

In certain range of score values the problem presents the tug-of-war alternative — boost of the reliability threshold (e.g. SEQUEST x-correlation value) rapidly decreases fraction of spectra that could be identified, while lowering it produces identifications from the "grey area", which are of dubious quality. Algorithmically, the only way out is to increase *information extraction* from tandem MS data. If we could somehow retrieve total information content of a given spectrum, its fate can be decided unambiguously depending on our capacity to learn from it. Such capability could be useful for a

* Presenting author

number of other interesting proteomics applications. The difficulties, of course, start with the definition of something as unusual as information content of peptide spectrum.

Recently we proposed Probability Profile Method (PPM) — classification algorithm that infers identities of the individual spectral peaks examining their spectral neighborhoods under the "microscope" of Bayesian statistics. PPM results have the form of probabilistic statements, like *peak number 123 is a b-ion with a 0.85 probability*. Efficient identification of "noble" b- and y-ion peaks dramatically simplifies construction of *de novo* tags (partial peptides) for a particular spectrum. Relatively simple algorithmic advances allowed us to build PPM-chain – tool for de novo protein tagging based on our methodology. During this study we have realized that traditional separation of MS computational algorithms into database search and de novo is very misleading. Our PPM-chain can be used in SEQUEST-emulation mode, taking full advantage of the known protein database, but at the same time has quite unique algorithmic capabilities, which include classical full-length de novo sequencing (it is not very good at the later task yet).

While capable of emulating SEQUEST our program works on entirely different mathematical and algorithmic principles. The laborious comparison between theoretical spectra and experimental spectra is the main CPU time consumer in database look-up algorithms, and correspondingly the performance typically scales linearly with the size of the search space, which grows exponentially in many situations (e.g., with the number of PTMs considered for each peptide). In de novo approach, almost all work is done up front, on the experimental spectra: peak labeling, finding of the tags, tag scoring. The need for the database comes very late in the process, involves very few candidate sequences and very simple procedures, which could be skipped all together for spectra with too little (no connectable peaks) or a lot (direct de novo identification) in terms of the informational content.

De novo identification also has inherent flexibility, which is reflected in the suppleness of its output. For a given spectrum and given specifications for de novo tag (e.g. 3 residues are set as a minimum length) PPM-chain has three possible outcomes: (1) "no tag" - no satisfactory de novo tag could be constructed for the spectrum; (2) "tag-no-match" – there are good de novo tags, but they do not conform to the available database; (3) "answer" – satisfactory de novo tag is found and mapped to a protein in the database. In contrast, database look-up programs return the best match with an attached score, which slowly decreases from the confidently identified spectra toward definite identification failures. In this case the bad quality of the match (e.g., because the database protein contains sequencing error) is hard to distinguish from the mediocre informational content of the spectrum (e.g., due to poor fragmentation). Such mix-up leads to all kinds of "grey area" situations, where valuable information - often indicative of unusual and interesting biological events - can be irrecoverably lost

We compared PPM-chain and SEQUEST using data sample obtained on the 54 ribosomal proteins of the *Rhodopseudomonas palustris* produced by Dr. Michael Strader at ORNL Center for Molecular and Cellular System. We have explored results of both programs for three spectral sets separated by SEQUEST X-correlation score: "high quality" (>3.2), "medium" (from 2.2 to 3.2) and "low quality" (<2.2) spectra. For the high quality subset "no tag" outcome was obtained only for 21 spectra (1.4%) and out of 1263 "answer" results SEQUEST identification was confirmed for 99.9% spectra. "Tag-no-match" outcome was observed for 216 cases (14%) and this fraction kept increasing in medium and low quality subsets: (~ 38% in both). The "answer" outcomes were still excellently aligned with SEQUEST ids (99% and 96% precision values, correspondingly). The fraction of "no tag" cases has grown sharply: 18% for medium and 57% for low confidence sets, reflecting the absence of the differentiating information in many spectra belonging there.

The result suggests an interesting speculation about possible sources of SEQUEST errors: it

is feasible that a large fraction of such errors is due to the absence of the underlying correct answer in the database. In such cases the returned match bound to be an incorrect one, but still may have relatively high X-correlation value. In our approach such spectra immediately become candidates for further study, such as Post Translational Modification (PTM) search or further de novo processing.
Summarizing, our testing indicates the following:

- Even with the existing technology (which certainly could be improved) reliable *de novo* tags can be constructed for a large majority of MS/MS spectra – and virtually for all high quality spectra.
- When de novo solution is compatible with the database, it is almost always the same as provided by SEQUEST. This conclusion confirms a high reliability of the SEQUEST identifications in the cases when the expected peptides are present in the protein database.
- There is a significant fraction of spectra (~33% for medium X-correlation values, ~50% low X-corr values), where the PPM-Chain finds good de novo tags not compatible with anything in the target database. Some of these tags definitely reflect complex and interesting biological phenomena, where PTMs and point mutations are blocking the possibility of finding the correct answers in the "plain vanilla" database searches.

For a future work we plan to apply PPM-chain for a comprehensive data extraction from the proteomics samples aiming at low abundance proteins as well as interesting biological facts, such as PTMs and mutated proteins. Our results strongly suggest that this approach will not only increase the output of useful information, but will also eliminate significant part of incorrect identification, further improving quality of the corresponding proteomics studies.

This work was funded in part by the US Department of Energy's Genomics:GTL program (www.doegenomestolife.org) under two projects, "Carbon Sequestration in *Synechococcus Sp.*: From Molecular Machines to Hierarchical Modeling" (www.genomes-to-life.org) and "Center for Molecular and Cellular Systems" (www.ornl.gov/GenomestoLife).

16

The Transcriptome of a Marine Cyanobacterium—Analysis Through Whole Genome Microarray Analyses

Brian Palenik[1]* (bpalenik@ucsd.edu), Ian Paulsen[2]* (ipaulsen@tigr.org), Bianca Brahamsha[1], Rob Herman[1], Katherine Kang[2], Ed Thomas[3], Jeri Timlin[3], and Dave Haaland[3]

[1]Scripps Institution of Oceanography, La Jolla, CA; [2]The Institute for Genomic Research, Rockville, MD; and [3]Sandia National Laboratories, Albuquerque, NM

Nitrogen and phosphorus abundance and type are thought to control photosynthesis and carbon sequestration in large areas of the world's oceans. Little is known about the regulation in cyanobacteria of nitrogen and phosphorus metabolism and their interaction with other environmental variables such as light and micronutrients. We are using a whole genome microarray of *Synechococcus* sp. WH8102 to examine these issues.

We have used whole genome microarrays in a number or experiments, initially to compare cells grown with nitrate and cells grown with ammonia. We found that 247 genes were down-regulated

* Presenting author

during growth under ammonia compared to nitrate. This included the NtcA transcriptional regulator known to control growth when ammonia is low and nitrogen sources other than ammonia are used. We also found that the use of a number of alternative nitrogen sources were down regulated (e.g. nitrate metabolism, cyanate transport, and urea metabolism). Some of these clearly have ntcA binding sites upstream although a complete comparison of the microarray results with ntcA binding site predictions are still in progress by Zhengchang Su and Ying Xu (UGA) as part of our GTL project. In addition, a number of genes associated with stress conditions are down regulated. These include glutathione peroxidase and a number of proteases and heat shock like proteins. This supports our hypothesis that growth under nitrate is actually more stressful than on ammonia due to the requirement of additional electron transport activity (and electron leakage) to reduce nitrate to nitrite to ammonia. Thus, we currently hope to analyze and model these results as a combination of NtcA regulation and stress response regulation. Interestingly a number of hypotheticals and conserved hypotheticals are down regulated, giving us an initial clue as to their possible functions in the cell.

In addition, we have also characterized phosphate limitation in WH8102 and made knockout mutants in a number of the two-component regulatory systems of the cell. We are also examining these phosphate limitation experiments with the wild type and mutant cells using whole genome microarray analyses. Because of its relatively small number of regulatory systems compared to many microbes, *Synechococcus* sp. WH8102 is an ideal model system for preparing a complete picture of the regulatory networks of an environmentally significant microbe.

Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-ACO4-94AL85000.

This work was funded by the U.S.Department of Energy's Genomics: GTL program (www.doegenomestolife.org) under project, "Carbon Sequestration in *Synechococcus Sp.*: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org) and by its Microbial Genome Project under "Transport and its regulation in marine microorganisms—a genomics-based approach."

17

DEB: a Data Entry and Browsing Tool for Entering and Linking *Synechococcus sp.* WH8102 Whole Genome Microarray Metadata from Multiple Data Sources

Arie Shoshani[1]* (Shoshani@lbl.gov), Victor Havin[1], Vijaya Natarajan[1], Tony Martino[2], Jerilyn A. Timlin[2], Katherine Kang[3], Ian Paulsen[3], Brian Palenik[4], and Thomas Naughton[5]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]Sandia National Laboratories, Albuquerque, NM; [3]The Institute for Genomic Research, Rockville, MD; [4]Scripps Institution of Oceanography, La Jolla, CA; and [5]Oak Ridge National Laboratory, Oak Ridge, TN

The process of generating and analyzing microarray data for *Synechococcus sp.* WH8102 whole genome in the Sandia-led GTL project involves three collaborators, where each generates metadata about their operation as well as data files. The *Synechococcus sp.* microbes are cultured in the Scripps Institution of Oceanography in San Diego, then the sample pool is sent to TIGR in Rockville, Maryland for microarray hybridization, 2-color scanning, and analysis. The scanned files and slides are then sent to Sandia Lab in Albuquerque, New Mexico for analysis and additional scanning with

a Hyperspectral Imaging instrument. Each of the institutes has an independent system for keeping track of metadata about their part of the operation and unfortunately these systems do not facilitate easy transfer of metadata details between institutions. This situation is typical of many biology projects, and it begs for a solution.

In this sub-project we set out to develop a single system where such metadata can be collected and linked in an orderly fashion. We developed a web-based Data Entry and Browsing (DEB) tool that can capture the metadata from experiments and laboratories and store them in a database in a computer searchable form. The key need is to have an easy-to-use intuitive system that integrates the metadata of all the related activities in this project. The design of the DEB tool is based on inputs and insights from the biologists on the project and as such contains features that a biologist will find useful. The interface design mimics the familiar laboratory notebook format. The system is built on top of the Oracle database system. The main concept of the interface design is to expose the biologist to a single "object" and its attributes at a time, and presenting objects as pages in a notebook that can be "turned", yet provide links between the objects in a simple intuitive fashion. An example of such a web-based screen is shown in the figure below.

The most powerful capability of the DEB system is that it is schema-driven, that is, all the interfaces to support all of the above features are generated automatically from the schema definition. Therefore, new metadata schemas can quickly be used to generate DEB interfaces as well as the underlying Oracle database for them. This feature makes this tool immediately applicable to new and/or changing databases. This allowed us to generate databases based on schemas designed by the biologists. Specifically, the scientists from the three sites have defined schemas for the Nucleotide Pool of microbes, for the Microarray Hybridization (based on the MIAME concepts), and the Hyperspectral Imaging and analysis system. The design included the ability to link these schemas and thus allow a researcher from any area of the project to extract metadata from the various parts of the experiment. For example the microarray hybridization schema has "probe_source" that links (points to) the "nucleotide_pool_id" in the Nucleotide Pool schema.

Data entry to the databases is done in two different modes: 1) on-line web-based data entry, and 2) automated data uploading from another database source. The on-line mode is used by the people who culture the Nucleotide Pools (Scripps), and the people running the Hyperspectral Imaging (Sandia). The automated data uploading is used for the Microarray Hybridization metadata (TIGR) because they have their own well-developed internal database system. The automated metadata loading is performed by dumping the metadata into simple formatted files (similar the spreadsheet output format) and have schema-driven tools for loading the data into the common database.

The main features of the DEB system are:

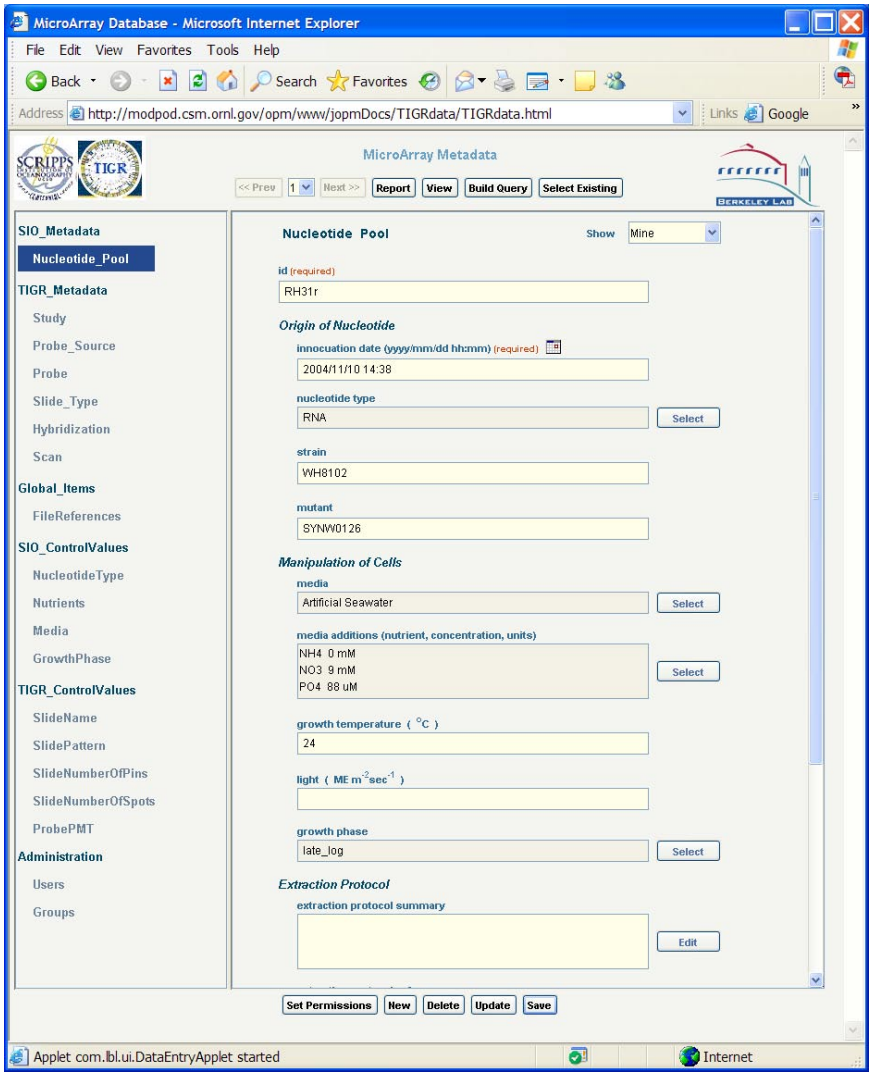
- It supports multiple inter-related object-classes, such as "experiment, materials, nucleotide-pool, samples, arrays, etc.
- For each object-class, it displays a page that mimics a notebook, with pages that can be "turned" (i.e. selected by previous-next, or by number)
- Objects can be linked to each other by simple connectors, such as a "sample" object linked to its "nucleotide-pool".
- Any file types (document, images, excel, etc.) can be uploaded to the system, and related to the metadata

* Presenting author

- Pages of the metadata can be printed for entry into a physical notebook – a requirement that makes sure the information is physically recorded
- Recording a new entry can be based on a previous entry, thus avoiding the re-entry of existing entries
- Security features to protect the metadata can be controlled by users and groups; each experiment or related objects can be assigned read, write, and delete permission to other users/groups.
- A query feature to search the metadata based on conditions on the attributes of the objects.

The importance of an easy-to-use system for capturing metadata in GTL cannot be overlooked, especially as an ever growing number of experiments are conducted and a large number of datasets are collected. The ability to quickly and automatically generate metadata systems from a schema description is essential for this evolving field with multiple sources of data gathered independently. DEB is currently running on LBNL's development server and at ORNL's GTL project operational server. While this system is designed for this project, its schema-driven architecture means that it can be applied to other GTL efforts.

Sandia is a multi-program laboratory operated by San-dia Corpo-



tion, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-ACO4-94AL85000.

18

Microarray Analysis using VxInsight and PAM

George S. Davidson*[1] (GSDAVID@sandia.gov), David Hanson[2], Shawn Martin[1], Margaret Werner-Washburne[2], and Mark D. Rintoul[1]

[1]Sandia National Laboratories, Albuquerque, NM and [2]University of New Mexico, Albuquerque, NM

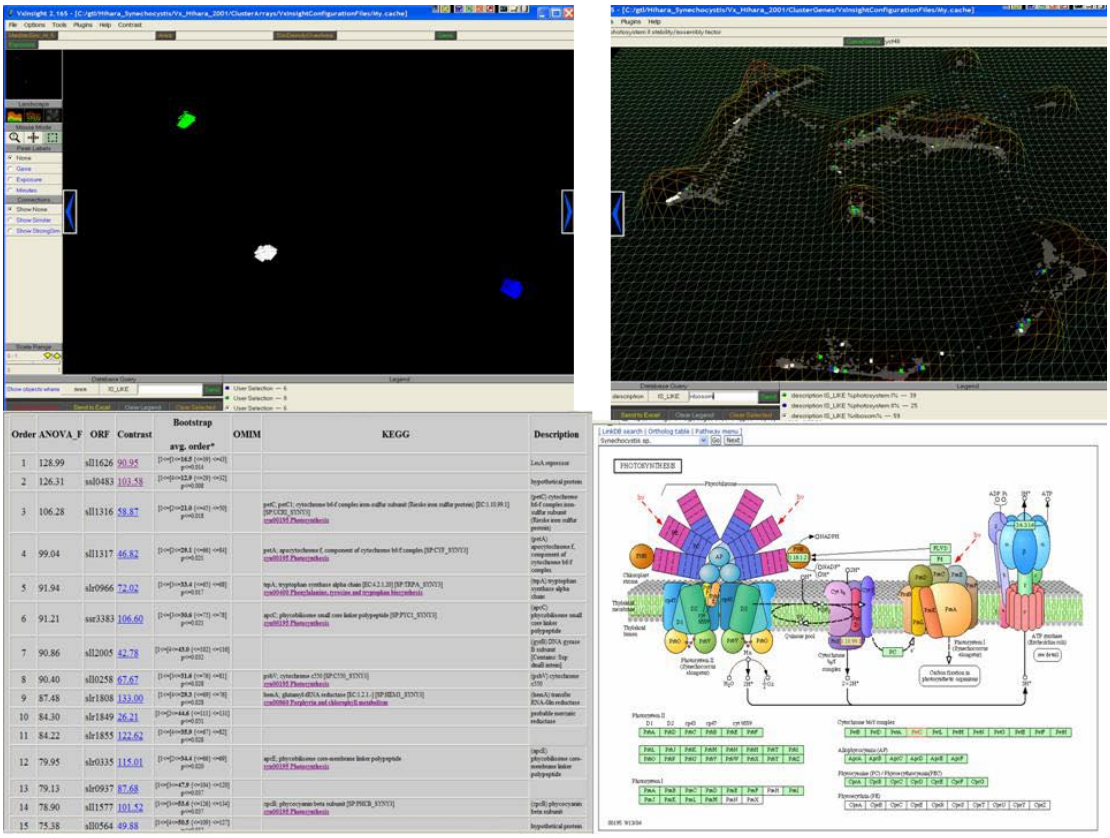
In 2001 Hihara *et al.* [1] published a series of microarray experiments describing gene expression changes in the cyanobacterium *Synechocystis sp.* PCC 6803 in response to acclimation from a low light level (20 μmol photons m^{-2} sec^{-1} to 300 μmol photons m^{-2} sec^{-1}). These data, which are publicly available from the KEGG database [2], have been reanalyzed using the VxInsight genome tools [3] from Sandia National Laboratories, and PAM [4] from Stanford. The analysis served as a test-bed for preparing the VxInsight tools to work with bacterial microarray data and associated databases. Here we present the results of clustering the individual experiments and the genes by co-expression. Interestingly, a number of experimental design issues are also raised. We examined lists of genes generated by PAM and by VxInsight that include significant differences in expression under the low light (LL) and the high light (HL) experimental conditions. We discuss the methodologies and the visual user interface linking these genes to online annotations and regulatory networks.

Hihara *et al.* measured total mRNA from HL conditions at 15 min, 1 hr, 6 hr, and 15 hr. These were compared to the mRNA levels measured under LL conditions, which served as control data. Expression levels were measured with CyanoCHIP version 0.8 from TaKaRa. These experiments revealed 84 ORFs with up regulated expression and 80 ORFs with down regulated expression after exposure to HL. Almost all of the photosystem I genes were immediately down regulated, while genes associated with photosystem II showed more complicated patterns. Both observations are consistent with the increasing PSII/PSI ratio in response to HL which involves an initial shift away from of PSI and the gradual construction of PSII (generally completed within 60 min.).

VxInsight found three groups of experiments, as shown in Figure 1, the first of which clearly reflects the stable, ongoing response to HL. The second captures the intermediate state when many of the PSII proteins are being synthesized, or have largely become available as the cells shift toward a higher metabolic plane. The third group contains a random mixture of arrays from time points throughout the experiment, which suggests that technical problems were confounding the measurements (something that is not uncommon in microarray experiments). We identified genes with significantly different expressions between late experimental conditions (first group) and the second group, which consisted of an equal number of measurements made at 15 minutes and at 60 minutes. VxInsight and PAM identified many of the same genes that Hihara *et al.* found; however the differences offer opportunities for deeper study. We present these genes with their scores and demonstrate the interactive analysis of these lists, including the use of KEGG pathways.

* Presenting author

Figure 1. The three clusters of arrays in the Hihara et al. experiment (top left), together with the gene clusters (top right). The top 15 genes relevant to the shift from HL to LL are listed on the bottom left, where the list contains links to KEGG networks, as shown on the bottom right.



Order	ANOVA_F	ORF	Contrast	Bootstrap avg. order*	OMIM	KEGG	Description
1	128.99	sll1626	90.95				LexA repressor
2	126.31	ssl0483	103.58				hypothetical protein
3	106.28	sll1316	58.87				
4	99.04	sll1317	46.82				
5	91.94	slr0966	72.02				
6	91.21	ssr3383	106.60				
7	90.86	sll2005	42.78				
8	90.40	sll0258	67.67				
9	87.48	slr1808	133.00				
10	84.30	slr1849	26.21				
11	84.22	slr1855	122.62				
12	79.95	slr0335	115.01				
13	79.13	slr0937	87.68				
14	78.90	sll1577	101.52				
15	75.38	sll0564	49.88				hypothetical protein

References

1. Hihara, Y., et al., *DNA microarray analysis of cyanobacterial gene expression during acclimation to high light.* The Plant Cell, 2001. **13**: p. 793-806.
2. Hihara, Y., http://www.genome.jp/kegg/expression/.
3. Davidson, G.S., *et al.*, *High throughput instruments, methods, and informatics for systems biology* (also to appear as: *Robust Methods for Microarray Analysis,* in *Genomics and Proteomics Engineering in Medicine and Biology*, IEEE/Wiley Press, Metin Akay, editor, in press*). 2003, Sandia National Laboratories SAND2003-4664: Albuquerque, New Mexico 87185.*
4. *Tibshirani, R., et al., Diagnosis of multiple cancer types by shrunken centroids of gene expression.* PNAS, 2002. **99**(10): p. 6567-6572.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. This work was supported by the U.S.Department of Energy's Genomics: GTL program (www.doegenomestolife.org) under project, "Carbon Sequestration in *Synechococcus sp.*: From Molecular Machines to Hierarchical Modeling,"(www.genomes-to-life.org).

19

Mapping of Biological Pathways and Networks across Microbial Genomes

F. Mao, V. Olman, Z. Su, P. Dam, and Ying Xu* (xyn@bmb.uga.edu)

University of Georgia, Athens, GA and Oak Ridge National Laboratory, Oak Ridge, TN

Homology exists beyond the individual gene level, and it could exist at the biological pathway and network level. There are a number of databases consisting of all experimentally validated and reliably predicted pathways/networks, providing a rich source of information for genome annotation and biological studies at a systems level. A key to effectively use such information is to identify orthologous genes accurately. However existing methods for mapping these known pathways and networks have serious limitations, greatly limiting the utility of such very useful information. Virtually all existing mapping methods are based on sequence similarity information, using tools such as reciprocal BLAST search or COG mapping. A fundamental problem with such methods is that sequence similarity information alone does NOT contain all the information needed to identify true orthologous genes!

We have recently developed a computational method and software, called P-MAP, for mapping a known pathway/network from one microbial organism to another by combining homology information and genomic structure information. The basic idea is that in microbes, genes working in the same pathway can generally be decomposed into a few operons or, in case of complex pathways/networks, regulons. Such information has not been effectively used in pathway mapping. When mapping known pathways, we first predict all the operons in a genome using our operon prediction program. The predictions are then validated through comparing microarray data mainly to check for consistency between gene expression patterns for genes predicted to be in the same operons or adjacent operons. Our evaluation has indicated that our prediction accuracy is close to 90%. With such information, we then map genes in a pathway template to the target genome that simultaneously gives relatively high sequence similarity between predicted orthologous gene pairs and has all the mapped genes grouped into a number of operons, preferably co-regulated operons based on the predicted *cis* regulatory elements and available microarray data. We have formulated the mapping problem as a linear integer programming (LIP) problem, and solved the problem using a commercial LIP solver, called COIN.

We have applied the P-MAP program to map known biological pathways in KEGG and MetaCyc to the cyanobacterial genomes and currently are mapping them to the *Shewanella oneidensis* MR-1 genome. Some of the mapping results could be found at http://csbl.bmb.uga.eddu/WH8102.

Acknowledgement: This project is supported by the U.S.Department of Energy's Genomics:GTL Program under project "Carbon Sequestration in *Synechococcus sp*: From Molecular Machines to Hierarchical Modeling" (http://www.genomes-to-life.org).

* Presenting author

20

Proteomic Analysis of the *Synechococcus* WH8102 CCM with Varying CO_2 Concentrations

Arlene Gonzales, Yooli K. Light, Zhaoduo Zhang, Michael D. Leavell, Rajat Sapra, Tahera Iqbal, Todd W. Lane, and Anthony Martino* (martino@sandia.gov)

Sandia National Laboratories, Livermore, CA

The genera *Synechococcus* and *Prochlorococcus* are oxygenic photoautotroph cyanobacteria. They are the most abundant picophytoplankton in the world's oceans where they form the foundation of the marine food web and are likely the largest contributors to primary production. Whole genome sequences are now available for a number of cyanobacteria including *Synechococcus* WH8102, *Prochlorococcus* MED4, and *Prochlorococcus* MIT9313. The sequences make it possible to use comparative analysis and high-throughput functional genomics and proteomics experiments to help better understand global diversity involved in carbon fixation.

Synechococcus WH8102's 2.4 Mb genome has yielded a number of interesting results regarding the carbon concentrating mechanism (CCM) in this organism. The carboxysome encoding operon in 8102 resembles that of β-proteobacteria rather than cyanobacteria. The operon most likely was acquired through horizontal gene transfer from phage. Carbonic anhydrase (CA) activity in the carboxysome shell protein csoS3 has been determined experimentally. Genome analysis indicates a putative β-CA and a ferripyochelin binding protein CA may also exist. Finally, transport of inorganic carbon in 8102 may occur through the low affinity CO_2 uptake genes ndhD4, ndhF4, and chpX. In *Prochlorococcus*, uptake genes have not been observed. Perhaps a unique transport mechanism exists in oceanic cyanobacteria.

We will present a high-throughput proteomic approach using mass spectrometry (MS), 2-hybrid analysis, and phage display to deconstruct components of the CCM and determine the effect of changing CO_2 levels in *Synechococcus* 8102. Protein expression levels of CCM components and protein-protein interactions within the carboxysome will be presented. Protein fractions were separated in to particulate and soluble fractions, and western blots of the fractions indicated rbcL and carboxysome shell proteins partitioned exclusively with the particulate fractions. Developing and fully mature carboxysomes were observed in the particulate fractions using electron microscopy. The two putative CAs partitioned separately in the particulate and soluble fractions. Changes in expression of specific proteins in cultures bubbled under different CO_2 levels were determined using 2D electrophoresis/MALDI-TOF MS. A synergistic whole proteome approach using capillary LC-MS/MS continues. Protein-protein interactions within the carboxysome have been determined using bacterial 2-hybrid techniques, and a number of pair wise interactions with be presented. Finally, rbcS-peptide interactions are being studied using phage display techniques.

21

Predicting Protein-Protein Interactions Using Signature Products with an Application to β-Strand Ordering

Shawn Martin[1] (smartin@sandia.gov), W. Michael Brown[1], Charlie Strauss[2], Mark D. Rintoul*[1], and Jean-Loup Faulon[3]

[1]Sandia National Laboratories, Albuquerque, NM; [2]Los Alamos National Laboratory, Los Alamos, NM; and [3]Sandia National Laboratories, Livermore, CA

As a part of the project entitled "Carbon Sequestration in *Synechococcus sp.*: From Molecular Machines to Hierarchical Modeling," we have developed a computational method for predicting protein-protein interactions from amino acid sequence and experimental data (Martin, Roe et al. 2004). This method is based on the use of symmetric tensor products of amino acid sequence fragments, which we call *signature products*. These products occur with different frequencies when considering interacting versus non-interacting protein pairs. We can therefore predict when a protein pair interacts by comparing the frequency of the signature products in that pair to the corresponding frequencies in protein pairs known to interact. Computationally, these comparisons are encoded into a Support Vector Machine (SVM) framework (Cristianini and Shawe-Taylor 2000), where the signature products are implemented using kernel functions. The final result is an automated classification system which can extrapolate from experimental results to complexes to entire proteomes, based only on experiment and primary sequence.

We have expended significant effort in benchmarking our method against competing techniques. In particular, we have compared the signature product method with methods based on products of InterPro signatures (Sprinzak and Margalit 2001; Mulder, Apweiler et al. 2003), concatenation of full-length amino acid sequences (Bock and Gough 2001), and a method combining multiple data sources (Jansen, Yu et al. 2003). In all cases our method performed as well as or better than the competing methods, as measured by 10-fold cross validation. We have applied our method to the prediction of protein-protein interactions in the case of yeast SH3 domains (Tong, Drees et al. 2002), where we achieved 80.7% accuracy, the full yeast proteome (69% accuracy), and *H. Pylori* (Rain, Selig et al. 2001) (83.4% accuracy). In addition to these results, which appear in (Martin, Roe et al. 2004), we have applied our method using COG networks (Tatusov, Koonin et al. 1997) to *Synechocystis sp.* (91% accuracy), and *Nostoc sp.* (69% accuracy).

Using our signature product approach, we have gone on to develop a method for ordering β-strands, which can in turn be used to improve the results of *ab initio* protein folding. The first step in our method is to train a signature product model for predicting β-strand interactions. This model was trained by extracting all β-strands from the Protein Data Bank (PDB) (Berman, Westbrook et al. 2000) using the dictionary of protein secondary structure (DPSS) method (Kabsch and Sander, 1983). After removing identical sequences from the ~20,000 structures available in the PDB, we obtained ~90,000 β-strands with more than 3 residues. Using the product signature method, we trained a β-strand interaction predictor which achieved 77% accuracy on a randomly selected training/test set combination with 80% of the β-strands in the training set and the remaining 20% of the β-strands in the test set. The next step in our method, as yet unimplemented, applies to individual proteins. For a given protein, we will apply our model to every possible pair of β-strands within the protein, and then consider every possible ordering of these strands. We will use the ordering which gives the highest

* Presenting author

average interaction score, as measured using our β-strand interaction predictor. We will validate our results by comparing our predicted ordering to the ordering actually present in the PDB.

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

22

In Vivo Observation of the Native Pigments in *Synechocystis sp*. PCC 6803 Using a New Hyperspectral Confocal Microscope

Michael B. Sinclair[1]* (mbsincl@sandia.gov), Jerilyn A. Timlin[1], David M. Haaland[1], Sawsan Hamad[2], and Wim F.J. Vermaas[2]

[1]Sandia National Laboratories, Albuquerque, NM and [2]Arizona State University, Tempe, AZ

We have developed a new hyperspectral confocal microscope that combines the attributes of high spatial resolution (<0.5 μm), high speed acquisition (>8 MB/s) and single photon sensitivity. The new instrument records the emission spectrum from 500 nm to 800 nm for each voxel within the 3-dimensional sample. The acquisition of full spectral information, when coupled with modern multivariate data analysis techniques allows for quantification of the contribution of each of the emitting components present within any voxel. To demonstrate the advantages of this approach, we have obtained and analyzed in vivo hyperspectral images of wild type *Synechocystis sp*. PCC 6803, as well as two mutant strains: chlL–, which is incapable of light-independent synthesis of chlorophyll, and PS1-less/chlL–, which in addition to being incapable of light-independent synthesis of chlorophyll does not assemble photosystem I. The raw emission spectra obtained from these specimens are quite complex, containing overlapping signatures from many pigments. Multivariate curve resolution analysis of the spectral images shows that each spectrum can be decomposed into independently varying contributions from phycocyanin, allophycocyanin, chlorophyll, a specific pool of "low-energy" chlorophyll associated with photosystem I, and protochlorophyllide. The relative contribution of each of these components varies from species to species in a manner consistent with expectations based on the genetic composition of the mutant strains. For example, we observe significant protochlorophyllide emission from the chlL– mutant which was grown in virtual darkness, while this emission is absent in the wild type. To our knowledge, this is the first ever demonstration of the coupling of rigorous deconvolution methods with hyperspectral confocal microscopy to reveal multiple overlapping native pigment emissions from in vivo specimens. We have also observed evidence for an inhomogeneous distribution of the emitting compounds within the cyanobacterium and are currently quantitatively exploring this inhomogeneity in the concentration distributions both within and between cells. This presentation will describe the design, construction and performance of the hyperspectral confocal microscope. Our results for *Synechocystis* will be described in detail.

Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-ACO4-94AL85000. This work was funded in part by the U.S.Department of Energy's Genomics: GTL program (www.doegenomestolife.org) under project, "Carbon Sequestration in *Synechococcus sp*: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org).

23

Connecting Temperature and Metabolic Rate to Population Growth Rates in Marine Picophytoplankton

Andrea Belgrano* (ab@ncgr.org) and Damian Gessler

National Center for Genome Resources, Santa Fe, NM

Photosynthetic picophytoplankton bacteria such as *Synechococcus* can contribute up to more then 50% of the total water column primary production, thus playing an important role in controlling the net flux of CO_2 between the atmosphere and the ocean, by the sequestration of carbon from the atmosphere via photosynthesis, and to global carbon cycling in marine systems.

We present how the rate of photosynthesis is regulated by changes in CO_2 concentration, irradiance and temperature, including allometric scaling theory and Rubisco activity as the primary catalyst, for the fixation of carbon by picophytoplankton.

To integrate physiological, biogeochemical, and environmental data in a single model, we integrate ¾-power scaling laws, along with irradiance, temperature, and nutrient uptake functions. Body mass scaling provides the theoretical and empirical model for constraints on the supply and use of energetic resources for picophytoplankton. We use a Boltzamann factor approach to capture the temperature-dependence of metabolism, and additionally introduce a Michaelis-Menten approach for nutrient uptake that includes cell quotas in a single growth model for *Synechococcus*.

We present evidence that a rise in oceanic temperature may reduce the relative contribution that picophytoplankton plays in the ocean as a carbon pathway. This highlights the importance of understanding shifts in the size composition of phytoplankton assemblages in relation to oceanic primary production of biomass, cell density, and nutrient status in the northwestern North Atlantic Ocean.

24

Deciphering Response Networks in Microbial Genomes through Data Mining and Computational Modeling

Z. Su[3], P. Dam[3], V. Olman[3], F. Mao[3], H. Wu[3], X. Chen[1], T. Jiang[1], B. Palenik[2], and Ying Xu[3]* (xyn@bmb.uga.edu)

[1]University of California, Riverside, CA; [2]Scripps Institution of Oceanography, San Diego, CA; and [3]University of Georgia, Athens, GA and Oak Ridge National Laboratory, Oak Ridge, TN

Deciphering of the "wiring diagrams" of biological networks (including metabolic, signaling and regulatory networks) represents a highly challenging problem, due to our lack of general understanding about the conceptual framework of how biomolecules work together as a system and an insufficient amount of experimental data. The majority of on-going computational research has been focusing on developing general methodologies for deriving "functionally equivalent" networks that are consistent with the limited experimental data such as microarray kinetics data, possibly leading to network topologies that are not biologically meaningful.

* Presenting author

We have been developing a computational framework, attempting to systematically derive network topologies that are most consistent with (a) information derived through mining genomic sequences and various genomic and proteomic data and (b) the kinetics data derived from microarray gene expression experiments. The framework consists of the following three key components:

1. **Identification of genes involved in a particular biological process:** To facilitate identification of genes possibly involved in a particular biological network, we first made genome-scale predictions of (a) gene functions, (b) operon structures and (c) *cis* regulatory elements at the genome scale. Gene function prediction is based on available genome annotation plus our own function prediction pipeline using additional information, including motif search and structure-based function prediction. Operons are predicted using our own program (see Section on operon/regulon predictions) *Cis* regulatory elements are predicted using our prediction program CUBIC, in conjunction with microarray data when available, through identification of conserved sequence motifs and similar gene expression patterns. Based on the initial identification of genes possibly involved in a particular biological network, we then refine/expend the gene candidate list through comparing to the information collected in (a), (b) and (c) described above.
2. **Prediction of interaction relationships among these candidate genes:** Currently we attempt to predict two types of interactions: (a) protein-protein interactions, both physical interactions and functional associations, and (b) protein-DNA interactions. Protein-protein interactions are predicted using homology search against protein interaction databases such as DIP & BIND and also based on prediction methods such as gene fusion/fission analysis, and phylogenetic profile analysis. We have developed our own methods for protein (transcription factors)-DNA interactions, based on both sequence and structural information. The sequence-based method is mainly based on (a) homology search against known protein-DNA complexes, (b) identification of self-regulation events, and (c) co-evolution information of transcription factors and operons they regulate. When the 3D structures of transcription factors are available, our method can accurately predict the binding affinity between the structure and its predicted DNA binding motifs, providing a highly effective tool for protein-DNA interaction prediction. Another piece of information for bio-molecular interactions comes from mapping known pathways from related organisms to the target genomes. Though not all such pathway mappings will provide complete pathway models in the target genome, the molecular interactions in the predicted pathways are useful and could be used for piecing together the "complete" network model in (3).
3. **Prediction of wiring diagrams through computational optimization:** We have developed two complementary methods for prediction of "complete" wiring diagrams of a target network, based on the predicted gene candidates and their (partial) interaction relationships and additional information. The first method connects the partially connected pieces predicted in (2) through mapping them to a genome-scale protein-protein interaction map we predicted in (2). The idea is to find the biologically most meaningful "paths" to connect the unconnected pieces (made of protein-protein and protein-DNA interactions). An algorithm has been developed for accomplishing this. In addition, we are currently developing a new algorithm that connects all the interaction components that are most consistent with the available microarray kinetics data, generalizing the current popular methods. By doing so, we can get wiring diagrams that are consistent with both molecular interaction information derived through data mining and microarray kinetics data.

We now describe two key procedures needed to implement the above computational prediction protocol because of the significance by their right.

Prediction of operons and regulons: We have recently developed a computational capability for prediction of operons in microbes, using multiple sources of information including (a) conserved gene neighborhoods across closely related organisms, (b) detected co-evolutionary information of genes, (c) functional relatedness of genes, (d) inter-genic distance information plus various types of other information. The overall prediction accuracy has reached 80% based on our test results on known operons in *E. coli*. We have applied this prediction program, in conjunction with available microarray data, to a number of genomes including *E. coli, Shewanella, Pyrococcus* and *Synechococcus*. The prediction procedure can be outlined as follows. We run the program to produce the initial operon candidate list and then we compare the predicted operons with available microarray data to check for consistency. Corrections will be made on the initial predictions if genes of the same operon exhibit significantly different expression patterns under any experimental condition, or genes from the neighboring operons have highly similar expression patterns under all known conditions and these operons are *very* close in the genomic sequence. In general about 5-10% of the original predictions are corrected based on the microarray data. We expect that the prediction accuracy could reach close to or even beyond 90% when sufficient microarray data is available. Based on the predicted operon structures, we then predicted regulons, based on available microarray data and genome-scale prediction of cis regulatory elements. The prediction procedure identifies operons that share similar expression patterns under the given experimental conditions and share conserved (predicted) binding motifs, and then clusters them into regulons. While this work is still in its early stage, we have identified a number of interesting regulons in the genomes we have applied prediction programs.

Pathway mapping: We have recently developed a computational method and software P-MAP for mapping a known pathway/network from one microbial organism to another by combining homology information and genomic structure information. The basic idea is that in microbes, genes working in the same pathway can generally be decomposed into a few operons or, in case of complex pathways/networks, regulons. Such information has not been effectively used in pathway mapping. When mapping known pathways, we first predict all the operons in a genome using our operon prediction program. The predictions are then validated through comparing microarray data mainly to check for consistency between gene expression patterns for genes predicted to be in the same operons or adjacent operons. Our evaluation has indicated that our prediction accuracy is close to 90%. With such information, we then map genes in a pathway template to the target genome that simultaneously gives relatively high sequence similarity between predicted orthologous gene pairs and has all the mapped genes grouped into a number of operons, preferably co-regulated operons based on the predicted *cis* regulatory elements and available microarray data. We have applied the P-MAP program to map known biological pathways in KEGG and MetaCyc to the cyanobacterial genomes and currently are mapping them to the *Shewanella oneidensis* MR-1 genome. Some of the mapping results could be found at http://csbl.bmg.uga.edu/WH8102.

Applications: We have applied this computational framework to predict the wiring diagrams of various response networks, which consists of signaling, regulatory and metabolic components. These include the carbon fixation, phosphorus assimilation and nitrogen assimilation networks in cyanobacterial genomes. Research is on going to apply the framework to *Shewanella oneidensis* MR-1.

Acknowledgement: This project is supported by the U.S.Department of Energy's Genomics:GTL Program under project "Carbon Sequestration in *Synechococcus sp*: From Molecular Machines to Hierarchical Modeling" (http://www.genomes-to-life.org)

* Presenting author

25

BiLab – A New Tool that Combines the Ease-of-Use of MatLab and the Power of Multiple Computational Biology Libraries

Al Geist* (gst@ornl.gov) and David Jung

Oak Ridge National Laboratory, Oak Ridge, TN

As part of the Genomics:GTL *Synechococcus*: From Molecular Machines to Hierarchical Modeling project, we are developing a new tool called BiLab that we hope will revolutionize computational biology the way the MatLab revolutionized numerical linear algebra. MatLab is widely used to do analysis, to develop new algorithms, and to teach students. MatLab is easy enough for new users and powerful enough for sophisticated users. Yet under the covers it is just a scripting language that provides easy access to the robust linear algebra functions in the LAPACK library. BiLab takes a similar approach except instead of only understanding matrices and doing linear algebra, BiLab understands biological objects such as DNA, proteins, and molecules and is able to manipulate them through any of the functions in a half-dozen standard computational biology libraries. And BiLab is able to display results in biologically relevant form, for example, a protein may be displayed as a molecule, a sequence alignment as stacked sequences.

We developed the BiLab scripting language to cater to different level of expertise in the user, from biologists who just want a quick way to use existing functions to bioinformatics programmers who want to write sophisticated programs in the BiLab scripting language. We have developed the tool to allow the easy addition of new biological objects and functions. Today BiLab provides access to all the functions in bioJava, bioPython, the text based NCBI tools, Jmol, JalView, and CDK. Developers can extend the scripting language to understand new biology. Thus the tool is designed to evolve with the Genomics: GTL program.

Like MatLab, data can be typed in manually or read in from files. BiLab understands the concept of remote biological databases and is able to dynamically load data from SwissProt, GenBank, FASTA, Protein Data Bank, EMBL, and other databases for analysis and study.

This presentation will describe how BiLab is built, how it can be extended, and hands-on demonstrations of the capabilities of the BiLab prototype.

Acknowledgement: This project is supported by the U.S.Department of Energy's Genomics:GTL Program under project "Carbon Sequestration in *Synechococcus sp*: From Molecular Machines to Hierarchical Modeling" (http://www.genomes-to-life.org)

26

Microbial Cell Modeling via Reacting/Diffusing Particles

Steve Plimpton* (sjplimp@sandia.gov) and Alex Slepoy

Sandia National Laboratories, Albuquerque, NM

We have developed a simulator called ChemCell [1] that tracks protein interactions within cells and can be used to model signaling, metabolic, or regulatory response. Cell features for microbial cells are represented realistically by triangulated membrane surfaces. Particles represent proteins, complexes, or other biomolecules of interest. They diffuse via 3d Brownian motion within the cytoplasm, or in 2d within membrane surfaces. When particles are near each other, they interact in accord with Monte Carlo rules to perform biochemical reactions, which can represent protein complex formation and dissociation events, ligand binding, etc. ChemCell is similar in spirit to MCell [2] and Smoldyn [3].

In this poster, we focus on the underlying algorithms used for reaction rules. We have recently developed a spatial version of the stochastic simulation algorithm (SSA) due to Gillespie [4] and discuss it's implementation in ChemCell. We compare it to alternative approaches including the original SSA and the interaction rules recently developed by Andrews and Bray [3]. We also highlight issues with various reaction/diffusion algorithms relevant to parallel implementation within ChemCell, with the eventual goal of enabling whole-cell models of realistic numbers of proteins and other biomolecules.

References:

1. S. J. Plimpton and A. Slepoy, SAND Report 2003-4509 (2003).
2. J. R. Stiles and T. M. Bartol, in Computational Neuroscience: Realistic modeling for experimentalists, edited by E. De Schutter, published by CRC Press, 87-127 (2001).
3. S. S. Andrews and D. Bray, Phys Biology, 1, 137-151 (2004).
4. D. T. Gillespie, J Comp Phys, 22, 403-434 (1976).

27

Modeling RuBisCO's Gating Mechanism Using Targeted Molecular Dynamics

Paul S. Crozier[1] (pscrozi@sandia.gov), Steven J. Plimpton[1], Mark D. Rintoul[1]*, Christian Burisch[2], and Jürgen Schlitter[2]

[1]Sandia National Laboratories, Albuquerque, NM and [2]Ruhr-Universität Bochum, Bochum, Germany

RuBisCO is the enzymatic bottleneck of carbon sequestration in *Synechococcus*, which is partly due to its catalysis of a competing oxygenase reaction that limits its specificity and efficiency. The binding niche residues are highly conserved across RuBisCO species, yet experimentally-measured specificities and carboxylation rates vary widely. The residues that make up the gate to the binding niche affect the gate's opening and closing rates, and in turn, RuBisCO specificities and reaction rates. We

* Presenting author

have performed molecular dynamics (MD) simulations of RuBisCO's gating mechanism to gain insight into how residue-level changes in RuBisCO's primary sequence affect enzyme performance.

Traditional MD is currently limited to the sub-microsecond timescale, but hardware and algorithm improvements continue to push the attainable timescales upward. We have recently developed several upgrades for our open-source parallel MD simulation package, LAMMPS (http://www.cs.sandia.gov/~sjplimp/lammps.html), which essentially double the algorithm's performance on typical biomolecular simulations. Performance enhancements have come through implementation of the rRESPA hierarchical time-stepping method and a high-performance tabulation algorithm for rapid evaluation of CPU-intensive coulombic interatomic forces. The LAMMPS simulation package was officially released as an open-source parallel MD code available for download on the first of September. Since then, it has been downloaded 1,075 times.

In addition to improving the traditional MD capabilities in LAMMPS, we have implemented advanced MD methods that allow simulation of events that occur on much longer timescales. One such method, targeted molecular dynamics (TMD), allows simulation of user-specified transition events, like the RuBisCO gating mechanism, by imposing a dynamic holonomic constraint on the macromolecular complex. TMD yields the free energy profile of the transition event, which is related to the rate of the transition event.

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

This project is supported by the U.S.Department of Energy's Genomics:GTL Program under project "Carbon Sequestration in *Synechococcus sp*: From Molecular Machines to Hierarchical Modeling" (http://www.genomes2life.org/). Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

28

Selection of Ligands by Panning of Phage Display Peptide Libraries Reveals Potential Partners for TPR Domain and rbcS in *Synechococcus* WH8102

Zhaoduo Zhang* (zzhang@sandia.gov), Arlene D. Gonzales, Todd W. Lane, and Anthony Martino

Sandia National Laboratories, Livermore, CA

One of the goals of functional genomics is the identification of reliable protein interaction partners. The oceanic cyanobacterium *Synechococcus* WH8102 is an abundant marine microorganism important to global CO_2 fixation. We have cloned, expressed and purified two TPR domains of a conserved hypothetical protein and the RuBiscO small subunit protein rbcS from *Synechococcus* WH8102. After immobilizing TPR domains and rbcS, selection of ligands were carried out by panning of two phage libraries displayed random peptides. Peptides specifically binding to TPR domain or rbcS were selected and enriched after three panning processes from a 7-mer and a 12-mer library. A sequence of three amino acids TPR or TPS forms a consensus peptide specific for TPR domains, and APL or APR forms a consensus specific for rbcS. The binding of clones to the target protein was further confirmed by ELISA assay. Peptides specifically binding to rbcS were found in carboxysome protein ccmK2, orfA, csoS3, csoS2 and rbcL, potential partners for rbcS.

* Presenting author